

COMPUTATIONAL AND EXPERIMENTAL FRAMEWORKS TO  
UNDERSTAND THE MECHANISM OF +1 AND -1 PROGRAMMED  
RIBOSOMAL FRAMESHIFTING

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

Pei-Yu Liao

February 2010

© 2010 Pei-Yu Liao

COMPUTATIONAL AND EXPERIMENTAL FRAMEWORKS TO  
UNDERSTAND THE MECHANISM OF +1 AND -1 PROGRAMMED  
RIBOSOMAL FRAMESHIFTING

Pei-Yu Liao, Ph. D.

Cornell University 2010

Programmed ribosomal frameshifting (PRF) is an extension of genetic decoding that allows a ribosome to produce inframe products and frameshift products from a single transcript. In +1 PRF, the ribosome moves one nucleotide to the 3'-end of the mRNA while in -1 PRF, the ribosome slips one nucleotide to the 5'-end of the mRNA during translation. Organisms from virus to prokaryotes to human have genes known to involve PRF. This dissertation presents the development of computational and experimental tools to systematically analyze the mechanism of +1 PRF and -1 PRF. The computational tools include: (1) a kinetic model for +1 PRF that reveals the synergistic effects of ribosome E-, P-, and A-sites on promoting +1 frameshift efficiency; (2) a mechanism-based bioinformatic program FSscan that identifies novel +1 frameshift cassettes in *yehP*, *pepP*, and *cheA* genes in *Escherichia coli*; and (3) a kinetic model for -1 PRF that predicts translation elongation steps significantly affecting -1 frameshift efficiency and the percentage of two types of -1 frameshift products. To confirm model predictions, a dual fluorescence reporter system is developed in *E. coli* and *Saccharomyces cerevisiae*. Additionally, -1 frameshift proteins are purified and analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry to obtain the percentage of the two types of -1 frameshift proteins. Using the reporter system in *E. coli*, the experimental results are consistent

with model predictions. The combination of computational and experimental works accelerates the investigation and expands the range and depth of the understanding. These tools can be further adapted to explore PRF in different organisms or to discover compounds altering PRF efficiency in a high-throughput manner. This work is an example of using systems biology approach to improve our understanding of a complex, but critically important, biological process.

## BIOGRAPHICAL SKETCH

The author was born in Taipei, Taiwan, a beautiful island in East Asia. She grew up in Taipei, where she completed her pre-undergraduate education. She likes to study science and is particularly interested in chemistry and biology. In Taipei First Girls' high school, she learned that a major in chemical engineering fitted her interest the best. Therefore, she applied to the Department of Chemical Engineering at National Taiwan University. She learned a broad range of engineering principles as well as biochemistry and molecular biology during her undergraduate program. Since then, she has decided to pursue advanced knowledge in biochemical engineering. In 2002, she entered the master program in the same department. She joined Prof. Shih-Yow Huang's biochemical engineering laboratory, where she conducted research in plant cell culture optimization. In August 2004, she came to the United States for the first time for her PhD study in the Department of Chemical and Biomolecular Engineering at Cornell University. She joined Prof. Kelvin H. Lee's research group, where she applied engineering principles to understand important biological systems. Staying in Ithaca, NY for three years, she knew a lot of good friends and she enjoyed the peaceful environment and the beautiful scenery. In August 2007, she moved with the group to University of Delaware at Newark, DE, where she knew new friends and was amazed by the opportunity for collaborations. In October 2007, she got married to her classmate in college, Wen-Shiue Young. Both of her and her husband's groupmates were witnesses in their civil marriage ceremony.

Dedicated to Chih-Chiang Liao and Ming-Chun Chien

## ACKNOWLEDGMENTS

I would sincerely thank my advisor Prof. Kelvin H. Lee. His great supporting and advising have allowed me to grow as a scientist. In my mind, he is the gold standard for being a good advisor and I hope by learning from him, I can get closer to this goal. I am grateful to my committee members, Prof. Michael L. Shuler, Prof. Matthew DeLisa, and Prof. Jeffrey D Varner at Cornell University. They have given me very helpful suggestions for my research. I would like to thank Prof. Jonathan D. Dinman at University of Maryland, College Park. He is always willing to share his expertise in ribosomal frameshifting field.

I am very fortunate to get help from a lot of individuals. I would like to thank Leila Choe, who taught me mass spectrometry and lab techniques. I am thankful for previous and current Lee group members: Erin Finehout, Chen Li, Kunal Aggarwal, Bob Kuczenski, Prateek Gupta, Mark D'Ascenzo, Dacheng Ren, Brenda Werner, Heather Roman, Jeff Swanberg, Gilda Shayan, Yong Choi, Anup Agarwal, Stephanie Hammond, and Jeff Foltz. Kunal taught me basic molecular biology skills and guided me to design my experiments. Bob taught me to write programs in Python. Prateek developed the the +1 frameshift kinetic model with me and helped me to troubleshoot in the lab. Yong run the protein samples by the mass spectrometry using multiple reaction monitoring. Everyone in our group has been very kind to me and I am so glad to be part of our group. I appreciate Dr. Alexey N. Petrov's help in developing the kinetic model of +1 frameshifting and Dr. Rasa Rakauskaite's assistance for performing the dual fluorescence reporter assay in yeast. I would like to thank two undergraduate students, Alice Tsai and Abhinav Rabindra Jian, who worked with me in 2007 and 2009. It has been a great experience to work with you.

My family and friends have been very supportive. I would like to thank my parents, sisters, Wan-Tsu and Pei-Ting, and brother Po-Yin. I always feel your care even if we are 8000 miles apart. I am truly grateful to my husband for his unwavering support. I would like to thank my roommates, Kai-Wen and Holly. I really enjoyed our time together. I would like to thank Yun-Wei, Lulu, Po-Hsun, Jason, Peter, Hsiu-Yu, Yi-Fan, and ChenPey for your friendships and great support.



## TABLE OF CONTENTS

<b>BIOGRAPHICAL SKETCH</b>	<b>iii</b>
<b>DEDICATION</b>	<b>iv</b>
<b>ACKNOWLEDGMENTS</b>	<b>v</b>
<b>TABLE OF CONTENTS</b>	<b>vii</b>
<b>LIST OF FIGURES</b>	<b>xiv</b>
<b>LIST OF TABLES</b>	<b>xviii</b>
<b>LIST OF ABBREVIATIONS</b>	<b>xx</b>
<b>CHAPTER 1: INTRODUCTION</b>	<b>1</b>
1.1 Background and motivation	1
1.2 Project goals	7
1.3 Scope of work	8
References	9
<b>CHAPTER 2: TRANSLATION ELONGATION</b>	<b>12</b>
2.1 Introduction	12
2.2 Transfer RNA	12
2.3 Ribosome	12
2.4 Translation elongation cycle	16
2.4.1 Steps in translation elongation	16
2.4.2 Kinetics of aa-tRNA selection	17
2.4.3 Kinetics of translocation	20
References	24

<b>CHAPER 3: DUAL FLUORESCENCE REPORTER SYSTEM</b>	<b>26</b>
3.1 Preface	26
3.2 Abstract	26
3.3 Introduction	27
3.4 Materials and methods	31
3.4.1 Stains and plasmids	31
3.4.2 Fluorescence assay	35
3.5 Results	37
3.5.1 Using the dual reporter system in <i>E. coli</i> to study the effect of Shine Dalgarno-like sequence on RF2 frameshifting	37
3.5.2 Using the dual fluorescence reporter in yeast to study the effect of anisomycin on -1 PRF	37
3.5.3 Using the dual fluorescence reporter in yeast to study the effect of ribosomal protein mutations on -1 PRF	39
3.6 Discussion	39
3.6.1 Dual fluorescence in <i>E. coli</i>	39
3.6.2 Dual fluorescence reporter in yeast <i>S. cerevisiae</i>	42
3.7 Supplementary data	43
3.8 Conclusion	44
References	45
 <b>CHAPTER 4: A NEW KINETIC MODEL REVEALS THE SYNERGISTIC EFFECT OF E-, P-, AND A- SITES ON +1 RIBOSOMAL FRAMESHIFTING</b>	 <b>48</b>
4.1 Preface	48
4.2 Abstract	48
4.3 Introduction	49

4.4 Kinetic model	51
4.5 Materials and methods	54
4.5.1 Computation of the kinetic model	54
4.5.2 Plasmids and bacterial strains	54
4.5.3 Fluorescence assay	55
4.5.4 Chi-square analysis	56
4.6 Results	56
4.6.1 Mathematical model.	56
4.6.2 Empirical studies	61
4.6.3 Parameter estimation	63
4.7 Discussion	65
4.7.1 Comparison of the three models	65
4.7.2 Role of the E-site	68
4.7.3 Role of the P-site	71
4.7.4 Role of the A-site	72
4.7.5 +1 PRF in eukaryotes	73
4.8 Supplementary data	74
4.9 Conclusion	80
4.10 Acknowledgments	80
References	81

<b>CHAPTER 5: FSSCAN: A MECHANISM-BASED PROGRAM TO IDENTIFY</b>	
<b>+1 RIBOSOMAL FRAMESHIFT HOT SPOTS</b>	<b>86</b>
5.1 Preface	86
5.2 Abstract	86
5.3 Introduction	87

5.4 FSscan algorithm	89
5.5 Materials and methods	92
5.5.1 Plasmids and bacterial strains	92
5.5.2 Fluorescence assay	93
5.5.3 Western analysis	95
5.5.4 Protein digestion	95
5.5.5 Liquid chromatography tandem mass spectrometry (LC-MS/MS)	95
5.6 Results	96
5.6.1 FSscan identifies a +1 frameshift hot spot in <i>prfB</i> gene	96
5.6.2 Analysis of 4132 protein coding sequences in the <i>E. coli</i> genome reveals additional potential +1 frameshift candidates	96
5.6.3 <i>In vivo</i> examination of +1 frameshift sequences agrees with the program	98
5.6.4 FSscan identifies <i>yehP</i> as a +1 frameshift candidate	98
5.7 Discussion	102
5.7.1 The scoring system	102
5.7.2 Analysis of six reading frames and pseudogenes	110
5.7.3 <i>yehP</i>	110
5.7.4 Other frameshift-prone sequences	112
5.7.5 FSscan as a bioinformatic program to search for novel +1 frameshift sequences	112
5.8 Supplementary data	113
5.9 Conclusion	120
5.10 Acknowledgments	120
References	121

## **CHAPTER 6: DIFFERENTIATING TWO TYPES OF THE FRAMESHIFT PROTEINS BY MASS SPECTROMETRY USING MULTIPLE REACTION**

<b>MONITORING</b>	<b>125</b>
6.1 Preface	125
6.2 Abstract	125
6.3 Introduction	125
6.4 Materials and methods	127
6.4.1 Plasmids and bacterial strains	127
6.4.2 Protein sample preparation	129
6.4.3 Mass spectrometry	129
6.5 Results	131
6.5.1 Correlation between peptide concentrations and peak area	131
6.5.2 Detection of two target peptides by nLC-ESI-MS/MS using MRM	133
6.6 Discussion	136
6.7 Conclusion	139
References	140

## **CHAPTER 7: KINETIC MODEL ANALYSIS OF THE EFFECT OF DIFFERENT ELONGATION STEPS ON -1 RIBOSOMAL FRAMESHIFTING**

	<b>142</b>
7.1 Preface	142
7.2 Abstract	142
7.3 Introduction	143
7.4 Kinetic model	147
7.5 Materials and methods	150
7.5.1 Computation of the kinetic model	150

7.5.2 Plasmids and bacterial strains	151
7.5.3 <i>In vivo</i> fluorescence assay	151
7.5.4 Protein purification and trypsin digestion	153
7.5.5 Mass spectrometry analysis	153
7.6 Results	153
7.6.1 Mathematical model	153
7.6.2 Experimental results	155
7.7 Discussion	161
7.8 Supplementary data	163
7.8.1 Mathematic model	163
7.8.2 Sensitivity analysis	169
7.9 Conclusion	169
References	172

<b>CHAPTER 8: FROM SNPS TO FUNCTIONAL POLYMORPHISM: THE INSIGHT INTO BIOTECHNOLOGY APPLICATIONS</b>	<b>176</b>
8.1 Preface	176
8.2 Abstract	176
8.3 Introduction	177
8.4 SNP analysis	180
8.4.1 SNP discovery	180
8.4.2 SNP detection	182
8.4.3 Study designs to relate a SNP to a phenotype	183
8.5 How SNPs lead to different phenotypes	184
8.5.1 DNA level: from DNA to RNA	186
8.5.2 RNA level: from RNA to protein	187

8.5.3 Protein level: from polypeptide formation to post-translation modification	190
8.6 Applications to the biotechnology industry	195
8.6.1 Breeding	195
8.6.2 Strain evolution	196
8.6.3 Biomolecule production	198
8.7 Challenges and the future	200
References	203
<b>CHAPTER 9: CONCLUSION AND FUTURE DIRECTIONS</b>	<b>212</b>
9.1 Summary	212
9.2 Future directions	213
9.2.1 Dual fluorescence reporter system in mammalian cells for therapeutic screening	213
9.2.2 Genome-wide scale PRF cassette identification	215
9.2.3 Investigating <i>yehP</i> , <i>pepP</i> , <i>nuoE</i> , and <i>cheA</i>	216
9.2.4 Applying FSscan to other genomes	218
9.2.5 Compositions of frameshift proteins in other -1 PRF cassettes	218
9.3 Conclusion	219
References	220

## LIST OF FIGURES

Figure 1.1. The genomic sequences of viruses that use programmed ribosomal frameshifting	4
Figure 1.2. Altering frameshift efficiency affects virus packaging	5
Figure 2.1. Cloverleaf secondary structure of a tRNA	13
Figure 2.2. A representative diagram for a ribosome structure	15
Figure 2.3. Translation elongation cycle	18
Figure 2.4. The hybrid site model	19
Figure 2.5. The kinetic model of aminoacyl-tRNA selection	21
Figure 2.6. The kinetic model of translocation	22
Figure 3.1. The genetic structure of the dual fluorescence reporter	30
Figure 3.2. The effect of Shine Dalgarno (SD)-like sequence on RF2 frameshifting in the <i>E. coli</i> system	38
Figure 3.3. The effect of anisomycin on -1 programmed ribosomal frameshifting in four genetic backgrounds	40
Figure 3.4. The effect of RPL3/TCM1 mutation on -1 programmed ribosomal frameshifting in four genetic backgrounds	41
Figure 4.1. The three kinetic models for +1 PRF in <i>E. coli</i>	52
Figure 4.2. The effect of P-site tRNA slippage (represented by $k_s$ ) and E-site tRNA release (represented by $k_r$ ) on FS% at fixed concentration of zero-frame cognate aa-tRNA ( $\text{cog.A}_0 = 1\%$ )	58
Figure 4.3. The effect of P-site tRNA slippage (represented by $k_s$ ) and the concentration of zero-frame cognate aa-tRNA ( $\text{cog.A}_0$ ) on FS% at fixed rate constant of E-site tRNA release ( $k_r = 100\text{ s}^{-1}$ )	59



Figure 4.4. The effect of the concentration of zero-frame cognate aa-tRNA (cog.A <sub>0</sub> ) and E-site tRNA release (represented by $k_r$ ) on FS% at fixed rate constant of P-site tRNA slippage ( $k_s = 5 \text{ s}^{-1}$ )	60
Figure 4.5. The effect of different E-site codon:anticodon interactions on frameshift efficiency	62
Figure 4.6. Data fit of frameshift efficiency for different codon:anticodon interactions in the E-site	66
Figure 4.7. (a) The correlation between FS% <sub>exp</sub> and the free energy change of the E-site codon:anticodon interactions. (b) The correlation between FS% <sub>exp</sub> and the apparent E-site stability obtained by free energy change of the E-site codon:anticodon interactions multiplied by modifying factors	67
Figure 5.1. The scoring system for FSscan program	91
Figure 5.2. FSscan identifies the +1 frameshift site in <i>prfB</i>	97
Figure 5.3. Maximum FSI in each of the 4132 <i>E. coli</i> protein coding sequences	99
Figure 5.4 Frameshift efficiency (FS%) for potential frameshift sequences identified by FSscan	100
Figure 5.5. Frameshift efficiency (FS%) for yehP6 and yehP7	103
Figure 5.6. (a) The nucleotide sequence design for yehP40, yehP41, and yehP4C. (b) Western blot for the cell lysate to detect the frameshift protein	104
Figure 5.7. Nucleotide and amino acid sequence for the YehP-EGFP frameshift protein in yehP41	105
Figure 5.8. Sequence conservation of the predicted frameshift cassette in <i>yehP</i>	108
Figure 5.S1. The correlation between log frameshift efficiency (FS%) and the stability difference in the P-site	117
Figure 5.S2. Frameshift efficiency (FS%) for selected P-site codons	118
Figure 5.S3. Tandem mass spectrum (MS/MS) of the peptide derived from the	

predicted frameshift site in <i>yehP</i>	119
Figure 6.1. Differentiating two types of frameshift products by mass spectrometry	128
Figure 6.2. Standard curves for HIV-1 target peptides	134
Figure 6.3. Correlation between the fraction of FS <sub>-1</sub> values calculated by concentration ratios (Eq.1) and peak area ratios (Eq.2)	135
Figure 6.4. Compositions of the frameshift proteins from four PRF cassettes by nLC-ESI-MS/MS using MRM	137
Figure 7.1. A mechanistic model of -1 programmed ribosomal frameshifting	145
Figure 7.2. The kinetic framework for -1 PRF	148
Figure 7.3. The effect of incomplete translocation (represented by $r_t$ ) on -1 PRF	156
Figure 7.4. The effect of the relocking step during translocation (represented by $r_4$ ) on -1 PRF	157
Figure 7.5. The effect of the slippage of P- and A-site tRNAs before peptidyl transfer (represented by $k_{pas2}$ ) on -1 PRF	158
Figure 7.6. The effect of the peptidyl transfer (represented by $k_{pt}$ ) on -1 PRF	159
Figure 7.7. Experimentally perturbing the system results in different levels of frameshift efficiency (FS% <sub>exp</sub> ) and the fraction of FS <sub>-1</sub>	160
Figure 7.S1. Sensitivity analysis using n-way analysis of variance (ANOVA)	170
Figure 8.1. An example of single nucleotide polymorphisms (SNPs) and the haplotype observed in six subjects	178
Figure 8.2. A possible framework for SNP applications in the biotechnology industry	181
Figure 8.3. An example of the effect of a SNP on DNA and transcriptional levels	188
Figure 8.4. Several examples of the effects of SNPs on mRNA and translational levels	191

Figure 8.5. An example of the effect of a SNP on the protein and post-transcriptional levels	193
Figure 9.1 A reporter system for a genome-scale screening for programmed ribosomal frameshift (PRF) cassette	217

## LIST OF TABLES

Table 3.1 Linker sequences for +1 PRF studies in Chapter 3	33
Table 3.2 Linker sequences containing -1 PRF signals in Chapter 3	34
Table 3.S1 Primers used in Chapter 3	44
Table 4.1 Three levels of +1 frameshift efficiency for different E-site codons	64
Table 4.S1 The rate constants for different steps at 20°C	76
Table 4.S2 The activation energy for different steps in the model and the fold change of the rate constants	77
Table 4.S3 The concentration of the components used in the model	78
Table 4.S4 Optimum parameter values for calculating $k_r$ obtained from data fitting	79
Table 5.1 Nucleotide sequences incorporated into the dual fluorescence reporter system for testing +1 frameshift efficiency <i>in vivo</i> in Chapter 5	94
Table 5.2 BLAST result for <i>yehP</i>	107
Table 5.S1 The stability of the base pairing in the ribosome P-site	115
Table 5.S2 Peptides detected by MRM and identified by ProteinPilot	116
Table 6.1 The nucleotide sequences incorporated into the dual fluorescence reporter for testing the compositions of frameshift protein in Chapter 6	130
Table 6.2 The list of target peptides and their MRM parameters	132
Table 7.1 Linker sequences and corresponding <i>E. coli</i> strains in Chapter 7	152
Table 7.S1 The rate constants for different steps during translocation at 37°C	166
Table 7.S2 The rate constants for different steps in aminoacyl-tRNA selection at 20°C	167
Table 7.S3 The activation energy for different steps in the model and the fold change of the rate constants	168
Table 7.S4 The concentration of the components used in the model	168

Table 7.S5 The list of target peptides and their MRM parameters	169
Table 8.1 Glossary of terms in single nucleotide polymorphism (SNP) studies	179
Table 8.2 Different types of genetic studies to connect SNPs to a phenotype	185

## LIST OF ABBREVIATIONS

aa-tRNA	aminoacyl-tRNA
AAV2	adeno-associated virus 2
ACS	1-aminocyclopropane-1-carboxylic acid synthase
AdoHcyase	S-adenosylhomocysteine hydrolase
ArgR	arginine repressor
CFTR	cystic fibrosis transmembrane-conductance receptor
CHO	Chinese hamster ovary
EGFP	enhanced green fluorescent protein
EST	expressed sequence tag
FA	formic acid
FS%	frameshift efficiency
GST	glutathione-S-transferase
GWA	genome-wide association
HIV	human immunodeficiency virus
MALDI-TOF/TOF	matrix assistance laser desorption/ionization tandem time-of-flight
miRNA	microRNA
MRM	multiple reaction monitoring
MS	mass spectrometry
nLC-ESI-MS/MS	nano-flow liquid chromatography electrospray tandem mass spectrometry
ORF	open reading frame
ODC	ornithine decarboxylase
PCR	polymer chain reaction

P-gp	P-glycoprotein
PhaC	Polyhydroxyalkanoate synthase
PHB	poly(3-hydroxybutyrate)
PRF	programmed ribosomal frameshifting
PXR	pregnane X receptor
RF2	release factor 2
SARS-CoV	coronavirus for severe acute respiratory syndrom
SD	Shine-Dalgarno
SNP	single nucleotide polymorphism
TIGR	tunable intergenic region
TPMT	thiopurine S-methyltransferase
UTR	untranslated region

## CHAPTER 1

### INTRODUCTION

#### ***1.1 Background and motivation***

During standard translation, genetic sequences are decoded in a successive, triplet manner. A sequence of triplets in mRNA not interrupted by a stop codon is defined as an open reading frame (ORF). Generally, a single ORF is expected to encode a single protein. However, the relation between a genome and its proteome is not linear. A single gene may produce multiple different proteins through different mechanisms such as alternative splicing of the mRNA transcript, varying translation start or stop sites, or frameshifting. All of these possibilities result in a proteome estimated to be an order of magnitude more complex than the genome [1].

Programmed ribosomal frameshifting (PRF) is a mechanism to expand the standard decoding. PRF is a coded shift in reading frame during translation of an mRNA transcript. In general, there are two types of PRF: in +1 PRF, the ribosome moves one nucleotide to the 3'-end of the mRNA while in -1 PRF, the ribosome slips one nucleotide to the 5'-end. Consequently, one transcript may yield two different protein products, an inframe product and a frameshift product. These products are likely to have distinct functions and contribute to the biological proteomic complexity. As of November 2009, organisms from virus to prokaryotes to human have genes known to involve PRF [2].

As mentioned above, +1 PRF has been observed in various organisms. In some cases, +1 PRF is known to be involved in gene regulation [3,4]. In *Escherichia coli*, the translation of *prfB* to produce release factor 2 (RF2) utilizes +1 PRF [3]. The reading



frame of the first 15% of this gene is determined by the start codon, while the remaining 85% of the mRNA is translated as a consequence of a +1 PRF event that involves bypassing the stop codon UGA in the frameshift site. Low RF2 levels result in inefficient recognition of the UGA codon, stimulating frameshift efficiency. High RF2 levels enhance the UGA decoding, reducing frameshift efficiency. The result is an autoregulatory feedback circuit in which RF2 levels control the production of RF2 via frameshifting. In mammalian cells, the expression of ornithine decarboxylase (ODC) antizyme involves +1 PRF [4]. Particularly, + 1 PRF provides an autoregulatory feedback loop between antizyme and polyamine levels. Antizyme degrades ODC, an enzyme that catalyzes the first and rate-limiting step of polyamine biosynthesis, decreasing polyamine production. When the level of polyamines rises, the +1 PRF efficiency increases, increasing the level of antizyme and thus reducing the abundance of ODC. The decrease in the ODC level will result in a lower polyamine level and consequently form a feedback loop.

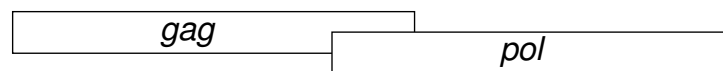
In *Saccharomyces cerevisiae*, retrotransposable elements, Ty1, Ty2, Ty3 and Ty4 [5-7], and three genes, *ABP140* [8], *EST3* [9], and *OAZ1* [10] employ +1 PRF. Ty1 uses +1 PRF to make a TYA-TYB protein required for transposition. The recoding site is CUU AGG C (where spaces separate zero frame codons), and the efficiency of the shift is greatly enhanced by the hungry zero-frame AGG codon in the A-site [5,11]. Ty3 is a retrotransposon that has overlapping genes *Gag3* and *Pol3*. *Gag3* encodes the structure protein for Ty3 virus-like particle and *POL3* encodes the enzymatic proteins required for transposition of Ty3 [12]. *POL3* overlaps with the last 38bp of *Gag3* in the +1 frame and does not have an independent ribosome entry site. The recoding site in Ty3 is GCG AGU U (where spaces separate the zero frame). Similar to Ty1, the AGU in the A-site of the Ty3 recoding site is a rare serine codon. The low availability

of cognate tRNA induces a pause, permitting more time for +1 PRF. +1 PRF permits 11% of the ribosome to enter *POL3* [6].

-1 PRF was first described in 1985 as the way of protein expression from the overlapping ORFs in retrovirus Rous Sarcoma Virus (RSV) [13]. Several important viruses, including human immunodeficiency virus type 1 (HIV-1) and the coronavirus responsible for severe acute respiratory syndrome (SARS), employ -1 PRF to produce proteins that are required for viral replication. In these viruses, the ORF encoding the viral structural protein (typically the Gag protein) is followed by an out of frame coding sequence for enzymatic proteins (typically Pro or Pol) (Figure 1.1). The enzymatic proteins are only expressed as a result of PRF with an efficiency of 1-40%, depending on the specific virus and the assay system. The frameshift efficiency determines the ratio of structural to enzymatic proteins available for virus particle assembly. Maintaining this ratio is important for virus particle morphogenesis because viruses require a large excess of structural components over the proteins with enzymatic activities (Figure 1.2). Previous studies showed that altering PRF efficiency affected the ratio of Gag to Gag-Pol protein synthesized and reduced viral titres [14-18]. Therefore, the PRF provides a unique target for antiviral agents. A reliable, low cost and easy-to-perform assay for PRF efficiency monitoring potentially permits a large scale screening of antiviral drugs.

In prokaryotes, -1 frameshifting has been documented for *dnaX* [19-21] and some insertion elements [22]. Certain bacteriophages employ -1 frameshift for the expression of tail genes [23,24]. In bacteriophage  $\lambda$ , a slippery sequence near the end of gene G, GGGAAAG, causes about 4% of -1 frameshift efficiency [25]. This ribosomal frameshifting results in producing longer proteins, gpGT. The standard

Viral genome



Viral proteins

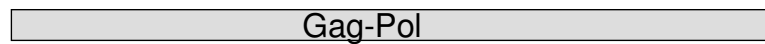


Figure 1.1. The genomic sequences of viruses that use programmed ribosomal frameshifting (adapted from [17]). The open reading frame (ORF) of the structure protein (e.g. *gag*) overlaps with the out of frame coding sequence of enzymatic proteins (e.g. *pol*). Programming ribosomal frameshifting results in a Gag-Pol fusion protein.

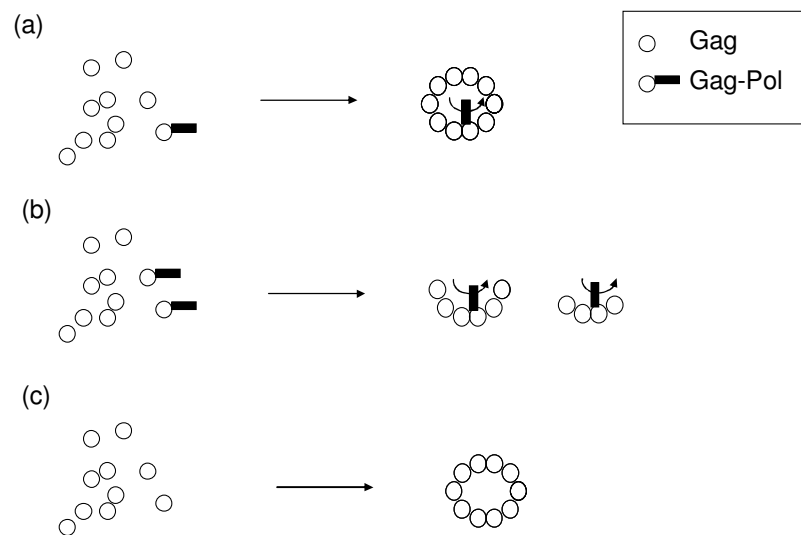


Figure 1.2. Altering frameshift efficiency affects virus packaging (adapted from [17]).  
 (a) The normal frameshift efficiency provides a correct ratio of Gag to Gag-Pol. (b) Increased frameshift efficiency results in incomplete viral particles. (c) Decreased frameshift efficiency results in particles without enzymatic proteins.

decoding product gpG and -1 frameshift product gpGT both participate in tail assembly.

PRF is suggested to be associated with several human disorders. Ribosomal +1 frameshifting within the human IL10 gene has been shown to produce a cryptic epitope recognized by cytotoxic T cells, which may play a role in precipitating autoimmunity [26]. In Alzheimer's disease, aberrant forms of  $\beta$ -amyloid precursor protein and ubiquitin have carboxyl-terminal amino acids encoded by a -1 reading frame of the mRNA. Therefore, Wills *et al.* [27] suggested a potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases. Supporting this hypothesis, in a polyglutamine-inducing neurodegenerative disorder spinocerebellar ataxia 3 (SCA3), the expanded CAG repeats in *MJD-1* transcript is prone to -1 frameshifting [28]. The result is the generation of *trans*-frame polyalanine-containing proteins. These polyalanine products may enhance polyglutamine-associated toxicity. Notably, the same study observed that anisomycin, which was previously shown to decrease -1 frameshifting [14,29], reduces the toxicity. Moreover, a prion protein was shown to cause different phenotypes by modulating frameshift efficiency in yeast *Saccharomyces cerevisiae* [30]. The prion [PSI<sup>+</sup>] is the amyloid conformation of the release factor 3 (eRF3) in yeast. [PSI<sup>+</sup>] can enhance antizyme production by promoting the +1 frameshifting required for antizyme expression. Because antizyme is a negative regulator of cellular polyamines, the increased level of antizyme by [PSI<sup>+</sup>] greatly reduces cellular polyamines level in yeast, resulting in distinct responses to environmental stress [30]. Recently, [PSI<sup>+</sup>] was also found to increase -1 PRF efficiency [31]. Taken together, PRF may involve in different types of diseases, either being a cause or a consequence of the disorder. The knowledge of PRF may thus provide insight into new therapeutic

strategies.

In summary, the ability of a single mRNA to encode more than one product would add to the information content of a genetic sequence, providing the molecule with a greater range of options. The discovery and characterization of new mechanism of PRF will further expand our understanding of translation control and the relation between ribosomes and *cis*-acting signals encoding in the mRNAs. Furthermore, knowledge of PRF may also shed light into novel therapeutic strategies for controlling virus or aberrant protein production.

### ***1.2 Project goals***

The overall objective of this project is to develop computational and experimental tools to understand the mechanism of +1 and -1 PRF. The main sub-goals of this work are:

1. Develop kinetic frameworks of +1 PRF and -1 PRF. The goal is to use the model to explain previous experimental observations and evaluate the significance of different parameters in the model.
2. Construct a reporter system in *Escherichia coli* and *Saccharomyces cerevisiae* to test PRF efficiency *in vivo*. The assay should be reliable, easy to perform, and low cost to permit a large scale application.
3. Perform a bioinformatic search for +1 PRF hot spots in the *E. coli* genome. The goal is based on sub-goal 1. The kinetic model of +1 PRF suggests several important features to induce +1 PRF. Therefore, searching the *E. coli* genome containing these features may allow identifying +1 PRF hot spots. The candidates sequences are further tested in the reporter system constructed in sub-goal 2.

### ***1.3 Scope of work***

The dissertation first provides background information for translation elongation, which is critical for understanding the mechanism of PRF (Chapter 2). Chapter 3 describes the development of a dual fluorescence reporter system to test PRF *in vivo*. This reporter system provides a tool to test the model predictions described in the following chapters. Chapter 4, adapted from [32], describes the kinetic model for +1 PRF and the experimental results validating the model predictions. Chapter 5, adapted from [33], illustrates the bioinformatic analysis for +1 PRF hot spots in the *E. coli* genome. Chapter 6 focuses on the method to analyze the composition of frameshift products using mass spectrometry. Chapter 7 presents a kinetic model for -1 PRF. Chapter 8 is a review of the effects of single nucleotide polymorphism (SNP) on phenotypes. A SNP may lead to different disease susceptibility, quantitative traits (e.g. height, weight, etc.) or drug responses in human. This review introduces how SNPs can be connected to a phenotype and the potential of SNPs analysis for biotechnological application. Finally, Chapter 9 summarizes the results and conclusions of the work and recommends areas of future direction.

## REFERENCES

1. Fields,S. (2001) Proteomics. proteomics in genomeland. *Science*, **291**, 1221-1224.
2. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2009) Recode-2: New design, new search tools, and many more genes. *Nucleic Acids Res.*, doi:10.1093/nar/gkp788.
3. Craigen,W.J. and Caskey,C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, **322**, 273-275.
4. Matsufuji,S., Matsufuji,T., Miyazaki,Y., Murakami,Y., Atkins,J.F., Gesteland,R.F. and Hayashi,S. (1995) Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell*, **80**, 51-60.
5. Belcourt,M.F. and Farabaugh,P.J. (1990) Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell*, **62**, 339-352.
6. Farabaugh,P.J., Zhao,H. and Vimaladithan,A. (1993) A novel programmed frameshift expresses the *POL3* gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell*, **74**, 93-103.
7. Janetzky,B. and Lehle,L. (1992) Ty4, a new retrotransposon from *Saccharomyces cerevisiae*, flanked by tau-elements. *J. Biol. Chem.*, **267**, 19798-19805.
8. Asakura,T., Sasaki,T., Nagano,F., Satoh,A., Obaishi,H., Nishioka,H., Imamura,H., Hotta,K., Tanaka,K., Nakanishi,H., et al. (1998) Isolation and characterization of a novel actin filament-binding protein from *Saccharomyces cerevisiae*. *Oncogene*, **16**, 121-130.
9. Morris,D.K. and Lundblad,V. (1997) Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr. Biol.*, **7**, 969-976.
10. Palanimurugan,R., Scheel,H., Hofmann,K. and Dohmen,R.J. (2004) Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J.*, **23**, 4857-4867.
11. Kawakami,K., Pande,S., Faiola,B., Moore,D.P., Boeke,J.D., Farabaugh,P.J., Strathern,J.N., Nakamura,Y. and Garfinkel,D.J. (1993) A rare tRNA-arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Genetics*, **135**, 309-320.
12. Hansen,L.J., Chalker,D.L., Orlinsky,K.J. and Sandmeyer,S.B. (1992) Ty3 *GAG3* and *POL3* genes encode the components of intracellular particles. *J. Virol.*, **66**, 1414-1424.



13. Jacks, T. and Varmus, H.E. (1985) Expression of the rous sarcoma virus pol gene by ribosomal frameshifting. *Science*, **230**, 1237-1242.
14. Dinman, J.D., Ruiz-Echevarria, M.J., Czaplinski, K. and Peltz, S.W. (1997) Peptidyl-transferase inhibitors have antiviral properties by altering programmed -1 ribosomal frameshifting efficiencies: Development of model systems. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 6606-6611.
15. Balasundaram, D., Dinman, J.D., Wickner, R.B., Tabor, C.W. and Tabor, H. (1994) Spermidine deficiency increases +1 ribosomal frameshifting efficiency and inhibits Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 172-176.
16. Dinman, J.D. and Wickner, R.B. (1992) Ribosomal frameshifting efficiency and gag/gag-pol ratio are critical for yeast M1 double-stranded RNA virus propagation. *J. Virol.*, **66**, 3669-3676.
17. Dinman, J.D., Ruiz-Echevarria, M.J. and Peltz, S.W. (1998) Translating old drugs into new treatments: Ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.*, **16**, 190-196.
18. Plant, E.P. and Dinman, J.D. (2008) The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front. Biosci.*, **13**, 4873-4881.
19. Blinkowa, A.L. and Walker, J.R. (1990) Programmed ribosomal frameshifting generates the *Escherichia coli* DNA polymerase III gamma subunit from within the tau subunit reading frame. *Nucleic Acids Res.*, **18**, 1725-1729.
20. Flower, A.M. and McHenry, C.S. (1990) The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 3713-3717.
21. Tsuchihashi, Z. and Kornberg, A. (1990) Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 2516-2520.
22. Chandler, M. and Fayet, O. (1993) Translational frameshifting in the control of transposition in bacteria. *Mol. Microbiol.*, **7**, 497-503.
23. Christie, G.E., Temple, L.M., Bartlett, B.A. and Goodwin, T.S. (2002) Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. *J. Bacteriol.*, **184**, 6522-6531.
24. Xu, J., Hendrix, R.W. and Duda, R.L. (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell*, **16**, 11-21.

25. Levin,M.E., Hendrix,R.W. and Casjens,S.R. (1993) A programmed translational frameshift is required for the synthesis of a bacteriophage lambda tail assembly protein. *J. Mol. Biol.*, **234**, 124-139.
26. Saulquin,X., Scotet,E., Trautmann,L., Peyrat,M.A., Halary,F., Bonneville,M. and Houssaint,E. (2002) +1 frameshifting as a novel mechanism to generate a cryptic cytotoxic T lymphocyte epitope derived from human interleukin 10. *J. Exp. Med.*, **195**, 353-358.
27. Wills,N.M. and Atkins,J.F. (2006) The potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases. *RNA*, **12**, 1149-1153.
28. Toulouse,A., Au-Yeung,F., Gaspar,C., Roussel,J., Dion,P. and Rouleau,G.A. (2005) Ribosomal frameshifting on MJD-1 transcripts with long CAG tracts. *Hum. Mol. Genet.*, **14**, 2649-2660.
29. Ioannou,M., Coutsogeorgopoulos,C. and Synetos,D. (1998) Kinetics of inhibition of rabbit reticulocyte peptidyltransferase by anisomycin and sparsomycin. *Mol. Pharmacol.*, **53**, 1089-1096.
30. Namy,O., Galopier,A., Martini,C., Matsufuji,S., Fabret,C. and Rousset,J.P. (2008) Epigenetic control of polyamines by the prion [PSI<sup>+</sup>]. *Nat. Cell Biol.*, **10**, 1069-1075.
31. Park,H.J., Park,S.J., Oh,D.B., Lee,S. and Kim,Y.G. (2009) Increased -1 ribosomal frameshifting efficiency by yeast prion-like phenotype [PSI<sup>+</sup>]. *FEBS Lett.*, **583**, 665-669.
32. Liao,P.Y., Gupta,P., Petrov,A.N., Dinman,J.D. and Lee,K.H. (2008) A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Res.*, **36**, 2619-2629.
33. Liao,P.Y., Choi,Y.S. and Lee,K.H. (2009) FSscan: A mechanism-based program to identify +1 ribosomal frameshift hotspots. *Nucleic Acids Res.*, doi:10.1093/nar/gkp796.

## CHAPTER 2

### TRANSLATION ELONGATION

#### ***2.1 Introduction***

Programmed ribosomal frameshifting (PRF) plays a crucial role in gene expression in both prokaryotes and eukaryotes (see reviews by Farabaugh [1] and Gesteland and Atkins [2]). The integrated model of PRF by Harger *et al.* proposed that +1 PRF and -1 PRF occurs at different phases of the elongation cycle [3]. Therefore, the mechanism of PRF should be understood within this context. This chapter first describes features of a tRNA and a ribosome structure, followed by a description of the mechanism of the translation elongation. A more detailed description of aminoacyl-tRNA (aa-tRNA) selection and translocation in the elongation are also presented.

#### ***2.2 Transfer RNA***

During protein synthesis, transfer RNAs (tRNAs) deliver amino acids to a growing polypeptide chain. Transfer RNAs consist of a single strand of RNA folded into a precise three-dimensional structure [4]. The tRNAs in bacteria and in the cytosol of eukaryotes have between 73 to 93 nucleotides. Most tRNA have a guanylate (pG) residue at the 5' end and all have the trinucleotide sequence CCA-3' at the 3' end. When drawn in two dimensions, the hydrogen-bonding pattern of all tRNAs forms a cloverleaf structure with four arms, or five arms for some longer tRNAs (Figure 2.1). Two of the arms of a tRNA are critical in protein synthesis: the acceptor arm can carry a specific amino acid and the anticodon arm contains the anticodon.

#### ***2.3 Ribosome***

Ribosomes are large ribonucleoprotein complexes that carry on translation in all cells.

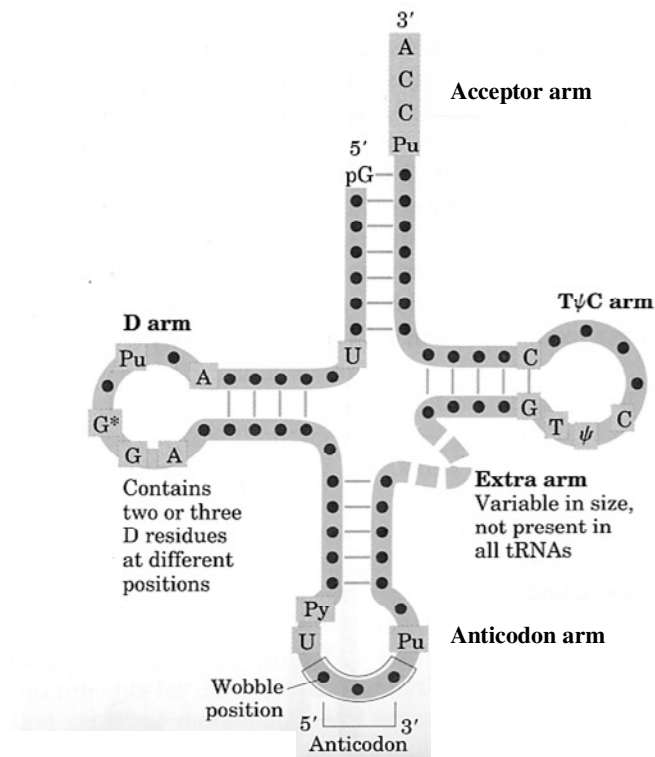


Figure 2.1. Cloverleaf secondary structure of a tRNA (adapted from [4]).

A ribosome is composed of two subunits. In prokaryotes, the subunits are designated 30S and 50S and together make up a 70S ribosome. The 50S subunit is composed of a 5S RNA subunit (consisting of 120 nucleotides), a 23S RNA subunit (3200 nucleotides) and 36 proteins. The 30S subunit has a 16S RNA subunit (1600 nucleotides) bound to 21 proteins [4]. A eukaryotic ribosome consists of a small (40S) and large (60S) subunit, together an 80S ribosome. The large subunit is composed of a 5S RNA (120 nucleotides), a 28S RNA (4700 nucleotides), a 5.8S subunit (160 nucleotides), and about 49 proteins. The 40S subunit has a 18S RNA (1900 nucleotides) and about 33 proteins [4].

A ribosome has three sites for tRNA binding: A-site, P-site, and E-site [5] (Figure 2.2). The ribosomal A-site is responsible for selecting cognate aa-tRNA and positioning of the aminoacyl moiety for the peptidyl transferase reaction. The A-site contacts the tRNA and mRNA with only four nucleotides of 16S rRNA [6]. The minimal interactions between the 30S A-site and the third-position codon and anticodon nucleotides permit the third-position wobble in the genetic code. The roles of the P-site include binding the initiator tRNA (a formyl-methionyl-tRNA in bacteria) during the initiation of protein synthesis and carrying the tRNA with a nascent polypeptide chain in a correct reading frame during elongation. On the small subunit, the P-site anticodon stem-loop (ASL) interacts with ten nucleotides of 16S rRNA [7] and the C-terminal tails of proteins S9 and S13, in addition to base pairing with its codon. In the 50S P-site, the minor groove of the P-site tRNA D stem rests on the minor groove of helix 69 of 23S rRNA, and the elbow of the tRNA contacts an extended  $\beta$ -hairpin loop of protein L5. A sharp kink was found in the mRNA between the A and P codons, which may allow simultaneous codon:anticodon pairing with both codons and contribute to maintaining the reading frame [8]. The E-site is believed to

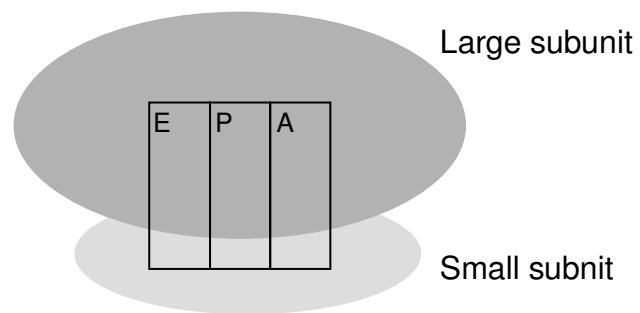


Figure 2.2. A representative diagram for a ribosome structure. A ribosome is composed of two subunits. E, P, and A denote the three tRNA binding sites in a ribosome.

provide a favorable free-energy change for movement of the deacylated tRNA out of the P-site [9,10]. In the recently solved *T. thermophilus* 70S ribosomal complexes [6,11], a non-cognate E-site tRNA interacts with 16S rRNA through a magnesium-mediated contact between tRNA phosphate 35 and 16S rRNA phosphate 1340. No codon:anticodon pairing has been observed in the E-site in these structures. However, whether the codon:anticodon interaction exists in the E-site is not known until the structure of an elongation complex containing a cognate E-site tRNA being resolved. The major interaction in the ribosome E-site is the binding of 3'-terminal adenosine on the tRNA (A76) to 23S rRNA, in which the tRNA must be deacylated and unmodified [12,13]. A crystal structure of *H. marismortui* 50S subunit revealed that A76 is positioned by extensive stacking and hydrogen-bonding with C2396 within 23S rRNA [14].

## ***2.4 Translation elongation cycle***

### **2.4.1 Steps in translation elongation**

During translation, the ribosome prolongs a polypeptide chain by adding a single amino acid in each elongation cycle (Figure 2.3). There are four basic steps in each cycle: (1) decoding: the ribosome recruits an aa-tRNA according the A-site codon. (2) accommodation: the ribosome accommodates the selected aa-tRNA. (3) peptidyl transfer: the peptidyl residue on the peptidyl tRNA is cleaved off and transferred to the aa-tRNA. As a result, the peptidyl-tRNA is now located at the A-site, extended by one amino acid. (4) translocation: the tRNA:mRNA complex moves by a codon length. In doing so, the peptidyl-tRNA enters the P-site, and the deacylated-tRNA enters the E-site. Translocation also brings a new codon into the A-site. With the selection of an aa-tRNA corresponding to this new codon, the ribosome enters into next elongation cycle. The allosteric three-site model for the ribosomal elongation cycle suggests a

reciprocal linkage between the E-site and A-site [15]. In this model, the E-site is occupied at the start of each cycle prior to aa-tRNA accommodation, and aa-tRNA binding promotes the release of the E-site tRNA, followed in turn by peptidyl transfer and translocation (Figure 2.3).

In the protein synthesis process, tRNAs move in the ribosome in order from A-, to P-, to E-sites. The hybrid-site model suggests that tRNA moves through the ribosome in an alternating fashion, with one end (the acceptor arm or the anticodon arm) fixed while the other is moving [10]. In this model, aa-tRNA accommodation has two states, A/T and A/A, while P-site tRNA is in P/P state (the notation before slash indicates the localization of the tRNA anticodon arm in the 30S subunit and that after slash indicates the localization of the tRNA acceptor arm in 50S subunit). Here, the T state refers to a specific site when the aa-tRNA is entering the ribosome. The two states of aa-tRNA imply that while the anticodon tip of the tRNA stays in the A-site in the 30S subunit, the acceptor arm of the tRNA moves from T- to A- site in the 50S subunit during accommodation (Figure 2.4). Similarly, during translocation, two tRNAs in the P/P and A/A first move their acceptor arms into the E- and P-sites in the 50S subunit, forming P/E and A/P. This step is followed by the movement of anticodon arms, which will form E and P/P states.

#### **2.4.2 Kinetics of aa-tRNA selection**

The ribosome recognizes aa-tRNA according to the match between anticodon and mRNA codon in the A-site. An aa-tRNA binds to the ribosome in a ternary complex with elongation factor EF-Tu and GTP in prokaryotes (eEF1 $\alpha$  in eukaryotes has functions analogous to EF-Tu). An elegant series of biochemical studies have produced a detailed kinetic model of aa-tRNA selection in the translation elongation



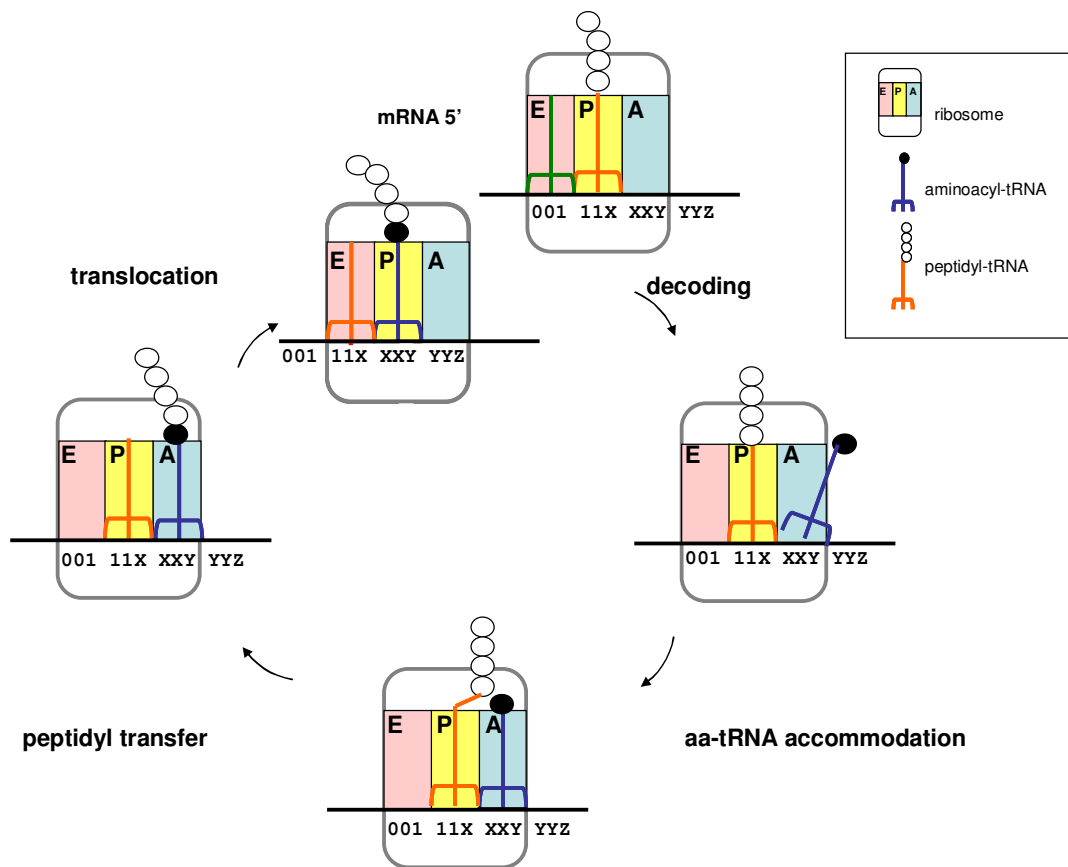


Figure 2.3. Translation elongation cycle. A cycle consists of four steps: decoding, aminoacyl-tRNA accommodation, peptidyl transfer, and translocation.

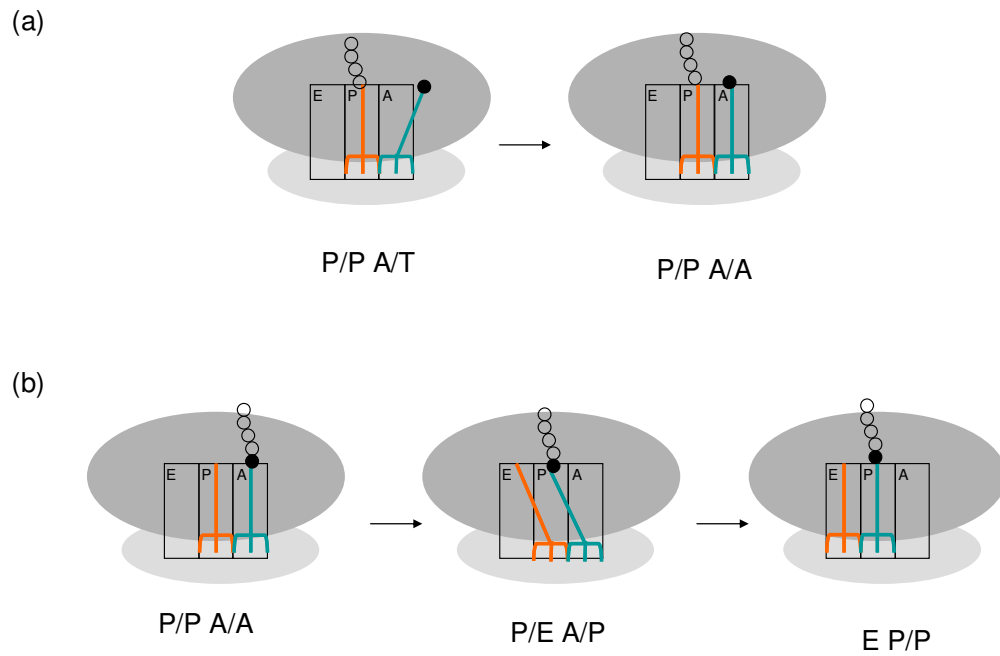


Figure 2.4. The hybrid site model. (a) aminoacyl-tRNA accommodation: while peptidyl tRNA stays in P/P, aminoacyl-tRNA moves from A/T to A/A. (b) translocation: deacylated-tRNA and peptidyl tRNA first move their acceptor arms to reach P/E and A/P, followed by moving their anticodon arms to reach E and P/P.

cycle [16]. In this model, fast initial binding of the ternary complex EF-Tu:aa-tRNA:GTP is followed by codon recognition. Codon recognition triggers EF-Tu GTPase activation, which leads to the GTP hydrolysis and dissociation of EF-Tu from the ribosome. Factor dissociation is followed by the spontaneous accommodation of the acceptor arm of the aa-tRNA into the A-site (Figure 2.5).

To ensure high fidelity during aa-tRNA selection, incorrect aa-tRNAs are rejected at two stages: initial selection of ternary complexes and proofreading of aa-tRNA. The ribosome contributes to the selection by enhancing the stabilities of correct codon-anticodon duplexes and accelerating the forward rates of GTPase activation and accommodation of a correct aa-tRNA. The rate of GTP hydrolysis for a cognate ternary complex is  $250\text{ s}^{-1}$ , while a near-cognate aa-tRNA with a C-A mismatch at the first position results in a GTP hydrolysis rate equal to  $0.4\text{ s}^{-1}$  [16]. Therefore, a much faster GTP hydrolysis of cognate compared to near-cognate substrate permits the selectivity of correct tRNAs. Similarly, a more restricted conformational space accessible to a cognate aa-tRNA increases the rate of the accommodation. In contrast, complete dissociation from the ribosome is favored for a near-cognate aa-tRNA in the absence of the interactions via EF-Tu and specific contacts to the closed form of the 30S subunit during proofreading.

### **2.4.3 Kinetics of translocation**

The last step of each elongation cycle is translocation, which involves the movement of two tRNAs and the mRNA on the ribosome (Figure 2.6). The translocation step is promoted by elongation factor G (EF-G) in prokaryotes [17,18] (eEF2 in eukaryotes has functions analogous to EF-G). The kinetic model of the translocation has been proposed previously [19]. The reaction sequence includes: binding of EF-G:GTP to

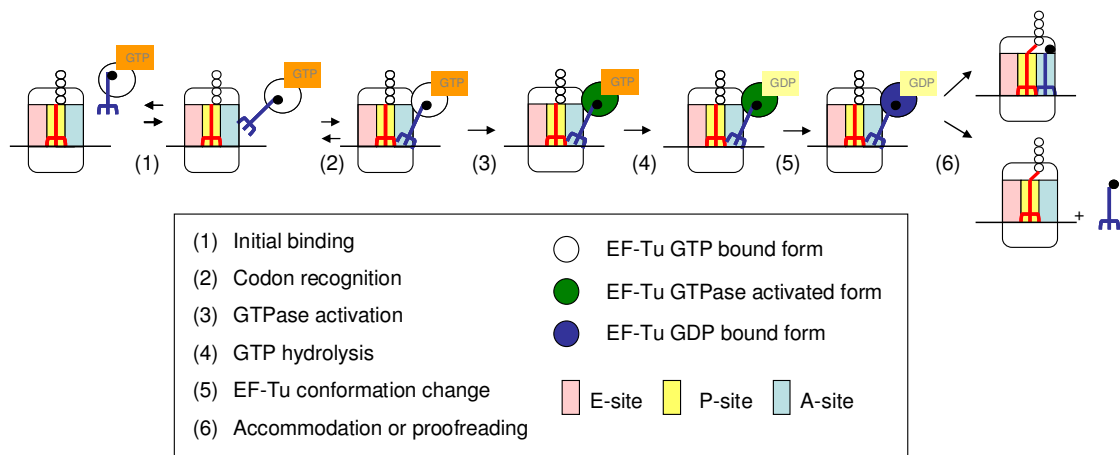


Figure 2.5. The kinetic model of aminoacyl-tRNA selection.

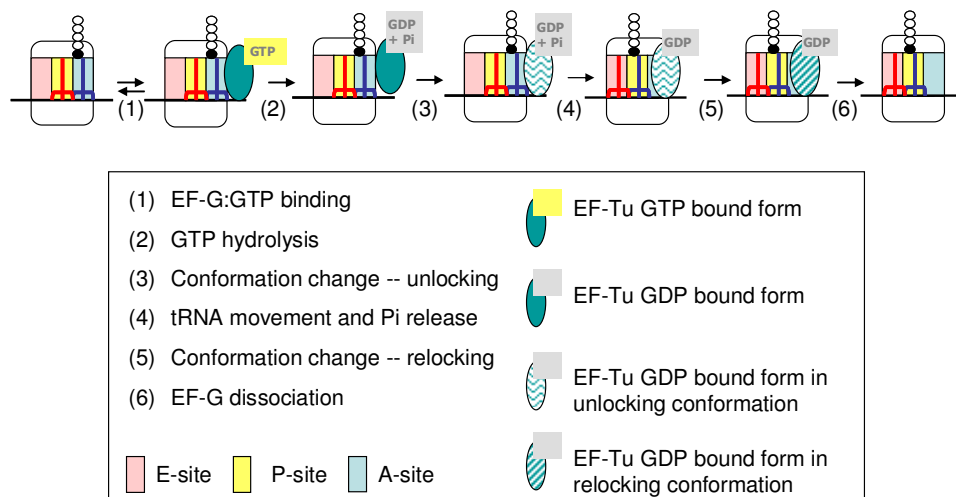


Figure 2.6. The kinetic model of translocation.

the pretranslocation complex, GTP hydrolysis, unlocking conformation change, tRNA movement with Pi release, relocking conformation change, and finally dissociation of EF-G:GDP and deacylated tRNA from the ribosome (Figure 2.6). Structurally, EF-G undergoes a reorientation and moves from a pretranslocation position outside the 30S A site to its posttranslocation position where domain 4 reaches into the 30S A-site, occupying the site of the anticodon arm of the A-site tRNA. This movement of EF-G may prevent backward tRNA movement. Therefore, EF-G may have two functions in the translocation: to unlock ribosome by coupling with GTP hydrolysis, and/or to bias tRNA-mRNA movement [20].

## REFERENCES

1. Farabaugh, P.J. (1996) Programmed translational frameshifting. *Annu. Rev. Genet.*, **30**, 507-528.
2. Gesteland, R.F. and Atkins, J.F. (1996) Recoding: Dynamic reprogramming of translation. *Annu. Rev. Biochem.*, **65**, 741-768.
3. Harger, J.W., Meskauskas, A. and Dinman, J.D. (2002) An "integrated model" of programmed ribosomal frameshifting. *Trends Biochem. Sci.*, **27**, 448-454.
4. Nelson, D.L. and Cox, M.M. (2005). *Principles of biochemistry*. 4th ed. W.H. Freeman and Co. NY. 2005.
5. Korostelev, A. and Noller, H.F. (2007) The ribosome in focus: New structures bring new insights. *Trends Biochem. Sci.*, **32**, 434-441.
6. Selmer, M., Dunham, C.M., Murphy, F.V., 4th, Weixlbaumer, A., Petry, S., Kelley, A.C., Weir, J.R. and Ramakrishnan, V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935-1942.
7. Guymon, R., Pomerantz, S.C., Crain, P.F. and McCloskey, J.A. (2006) Influence of phylogeny on posttranscriptional modification of rRNA in thermophilic prokaryotes: The complete modification map of 16S rRNA of *Thermus thermophilus*. *Biochemistry*, **45**, 4888-4899.
8. Yusupova, G.Z., Yusupov, M.M., Cate, J.H. and Noller, H.F. (2001) The path of messenger RNA through the ribosome. *Cell*, **106**, 233-241.
9. Lill, R., Robertson, J.M. and Wintermeyer, W. (1989) Binding of the 3' terminus of tRNA to 23S rRNA in the ribosomal exit site actively promotes translocation. *EMBO J.*, **8**, 3933-3938.
10. Moazed, D. and Noller, H.F. (1989) Intermediate states in the movement of transfer RNA in the ribosome. *Nature*, **342**, 142-148.
11. Korostelev, A., Trakhanov, S., Laurberg, M. and Noller, H.F. (2006) Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, **126**, 1065-1077.
12. Feinberg, J.S. and Joseph, S. (2001) Identification of molecular interactions between P-site tRNA and the ribosome essential for translocation. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 11120-11125.
13. R. Lill *et al.*, Specific recognition of the 3'-terminal adenosine of tRNA<sup>Phe</sup> in the exit site of *Escherichia coli* ribosomes, *J. Mol. Biol.* **203** (1988), pp. 699–705.

14. Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2003) Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit. *RNA*, **9**, 1345-1352.
15. Nierhaus, K.H. (1990) The allosteric three-site model for the ribosomal elongation cycle: Features and future. *Biochemistry*, **29**, 4997-5008.
16. Pape, T., Wintermeyer, W. and Rodnina, M. (1999) Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. *EMBO J.*, **18**, 3800-3807.
17. Rodnina, M.V., Savelsbergh, A., Katunin, V.I. and Wintermeyer, W. (1997) Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome. *Nature*, **385**, 37-41.
18. Katunin, V.I., Savelsbergh, A., Rodnina, M.V. and Wintermeyer, W. (2002) Coupling of GTP hydrolysis by elongation factor G to translocation and factor recycling on the ribosome. *Biochemistry*, **41**, 12806-12812.
19. Wintermeyer, W., Peske, F., Beringer, M., Gromadski, K.B., Savelsbergh, A. and Rodnina, M.V. (2004) Mechanisms of elongation on the ribosome: Dynamics of a macromolecular machine. *Biochem. Soc. Trans.*, **32**, 733-737.
20. Savelsbergh, A., Katunin, V.I., Mohr, D., Peske, F., Rodnina, M.V. and Wintermeyer, W. (2003) An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell*, **11**, 1517-1523.



## CHAPTER 3

### DUAL FLUORESCENCE REPORTER SYSTEM

#### **3.1 Preface**

An *in vivo* reporter system is essential to experimentally study different effects on programmed ribosomal frameshifting. This chapter describes the development of a dual fluorescence reporter system in *Escherichia coli*. The reporter system was later transferred into yeast *Saccharomyces cerevisiae* to test several PRF signals in eukaryotic cells.

#### **3.2 Abstract**

Programmed ribosomal frameshifting (PRF) is the process by which ribosomes produce two different polypeptides from the same mRNA. A new *in vivo* dual fluorescence reporter system is developed to study PRF in *Escherichia coli* and yeast *Saccharomyces cerevisiae*. Frameshift sites are inserted between two fluorescence reporter genes, monomeric DsRed and EGFP, contained in an *E. coli* expression vector or a yeast shuttle vector. The red and green fluorescence for different strains are directly measured by a microwell plate reader. The system allows an easy comparison of frameshift efficiency for different recoding sites with the normalized fluorescence ratio. By using the system in *E. coli*, the integrity and the position of the stimulatory signal (Shine-Dalgarno-like sequence) are shown to affect +1 PRF efficiency significantly. In addition, PRF signals from HIV-1 group M, group O, and yeast L-A virus were tested in the yeast system. The dual-fluorescence reporter system has potential as a high-throughput and non-invasive *in vivo* assay for PRF studies.

### 3.3 Introduction

During protein synthesis, ribosomes translate the nucleotide sequence of an mRNA molecule into the amino acid sequence of a protein. Normally, ribosomes read mRNAs in successive, adjacent three nucleotide (triplet) codons. The chance for the ribosome to switch its reading frame during the translation occurs at a very low frequency, about  $10^{-4}$ - $10^{-5}$  per codon [1]. Programmed ribosomal frameshifting (PRF) is a coded shift in reading frame during translation of an mRNA transcript, in which the frameshift can occur at rates from 1000- to 10,000- fold higher than nonprogrammed sites [2]. Consequently, one transcript may yield two different protein products, an inframe product and a frameshift product. PRF has been observed to occur in various organisms including prokaryotes and eukaryotes [3,4]. Several important viruses including human immunodeficiency virus type 1 (HIV-1) and the coronavirus for severe acute respiratory syndrome (SARS-CoV), employ -1 PRF to synthesize the precursor of enzymes for their replication [5,6]. Previous studies have demonstrated that altering -1 PRF efficiency may damage the viral replication (see Chapter 1 and the review by Dinman *et al.* [7]). Therefore, a system monitoring the change in PRF can provide a platform for antiviral drug screening.

A widely used reporter system for studying frameshifting events is based on the  $\beta$ -galactosidase assay. The frameshift sites are followed by *lacZ* gene in a different reading frame. This sequence is designed to ensure the production of the *lacZ* gene product,  $\beta$ -galactosidase, to be dependent upon PRF events. Frameshift efficiency is measured by determining the ratio of  $\beta$ -galactosidase activity produced from a construct requiring PRF to express *lacZ* to that of a construct in which the *lacZ* is in frame [8-10]. The enzymatic-based assay is highly quantitative. However, this reporter system does not measure the ratio between the zero-frame translation and PRF events

within the same strain. Thus, complicating factors may arise, such as the need to normalize for cell number, protein concentration, mRNA abundance, and differential translational efficiencies of the PRF reporter and zero-frame reporter mRNAs.

Bicistronic reporter systems present a strategy to internally control for the variability between different mutants. In the studies by Baranov *et al.* [11] and Hansen *et al.* [12], test frameshift cassettes were inserted between glutathione-S-transferase (GST) and *malE* genes and the *malE* gene was made in the +1 frame. Measurements of frameshift efficiency were done by image analysis of SDS gels. Frameshift efficiency was estimated from the amount of frameshift product divided by the total protein synthesized from the *GST-malE* reporter (frameshift product + non-frameshift product). Grentzmann *et al.* [13] developed a dual-luciferase reporter system for studying recoding signals *in vitro*. The dual-luciferase assay simultaneously measured the luminescence of both the *Renilla* and firefly luciferase enzymes synthesized from a single bicistronic mRNA. The two genes are separated by a functional PRF signal and the downstream firefly gene is placed into the -1 or +1 frame relative to the upstream *Renilla* gene. The relative luciferase expression of firefly to *Renilla* is normalized by a zero-frame control plasmid that lacks a frameshift signal and has firefly luciferase gene in the zero frame. Harger *et al.* [14] further applied this system in yeast for measurement of Ty1 and Ty3 directed +1 frameshifting. However, this analysis requires cell lysis, and is relatively expensive and labor consuming.

Using green fluorescent protein (GFP) of the jellyfish *Aequorea victoria* as a reporter requires neither substrates nor cofactors due to the intrinsically fluorescent nature of the protein. Since this gene was first cloned in *E. coli* and *Caenorhabditis elegans* in 1992 [15], efforts have been put into engineering GFP to produce variants with

different color, enhanced folding efficiency, increased stability, or altered oligomerization (see a review by Tsien [16]). The enhanced GFP (EGFP) contains Phe64L and Ser65Thr mutations in the protein and is one of the brightest variant of GFP [17]. More recently, fluorescent proteins from other species have been identified and isolated [18,19]. The various options of fluorescent proteins provide tools for multicolor labeling, fluorescence resonance energy transfer, or a multicolor reporter assay [20-22]. Choe *et al.* used a dual fluorescence reporter for a high-throughput clone characterization assay [21]. This dual reporter system was designed such that a successful insertion of the foreign DNA caused the loss of DsRed fluorescence without interrupting GFP fluorescence. Combined with cell sorting, this approach provided rapid screening for isolating clones with successful recombination. To test the effect of tunable intergenic regions (TIGRs), Pfleger *et al.* incorporate various TIGRs between DsRed and EGFP under the control of the same promoter [22]. By measuring red and green fluorescence levels, this study observed that TIGRs can vary the relative expression of two reporters in the same operon over a 100-fold range.

In the present study, a dual fluorescence reporter system for studying PRF is developed in *E. coli* and *S. cerevisiae*. PRF signals are inserted between the coding sequences of DsRed and EGFP (Figure 3.1). In the test strain, the expression of the first fluorescent reporter (DsRed in the example of Figure 3.1) depends on the translation efficiency, while the expression of the second reporter depends on both translation and PRF efficiencies. Therefore, the system enables comparisons of frameshift efficiency for different recoding sites with the normalized fluorescence ratio. Because the detection of fluorescence requires neither cell lysis nor enzymatic assays, the dual fluorescence reporter system is a non-invasive and cost-effective assay to monitor PRF efficiency.

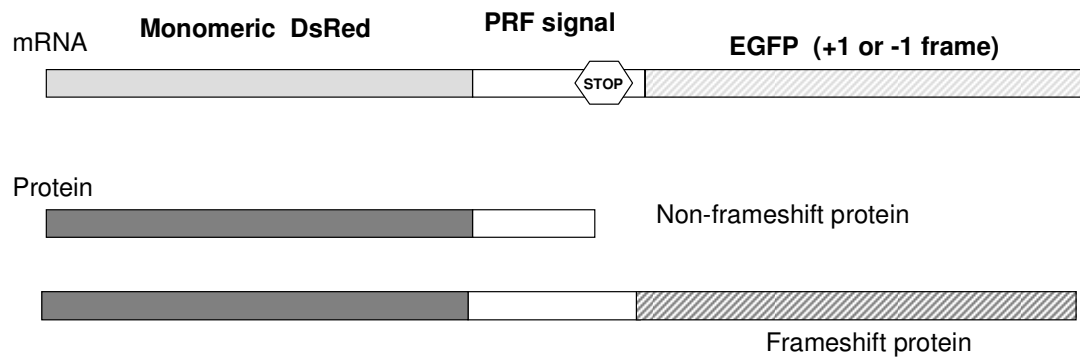


Figure 3.1. The genetic structure of the dual fluorescence reporter. The coding sequence of monomeric DsRed is upstream of an out of frame *egfp*. The linker sequence contains a stop codon in-frame with the DsRed coding sequence. As a result, the reporter system expresses DsRed as non-frameshift proteins and DsRed-EGFP as frameshift proteins.

### 3.4 Materials and methods

#### 3.4.1 Stains and plasmids

*Escherichia coli* XL1 blue MRF' (Stratagene, La Jolla, CA) was used in the *E. coli* system. All primers for PCR amplification and mutagenesis in this study were obtained from Integrated DNA technologies (Coralville, IA) and listed in Table S3.1 in Supplementary Data. The gene sequence of monomeric DsRed was PCR amplified by using two primers, Hind-mRFP and mRFP-SalI, and pmRFP [23] as a template. The PCR product was cloned into HindIII/SalI sites in pEGFP vector (Clontech, Mountain View, CA) to create pRG3 plasmid, which can express DsRed-EGFP fusion protein. Different linker sequences were made from complementary oligonucleotides (Integrated DNA technologies) and were cloned between SalI and BamHI sites between the coding sequences of DsRed and EGFP in the pRG3 plasmid. The linker sequence for a control kept both DsRed and EGFP coding sequences in frame. In a test strain, the linker sequences containing mutated release factor 2 (RF2) frameshift sites rendered downstream *egfp* in the +1 frame. The control strain expressed only the DsRed-EGFP fusion protein from the reporter. The test strains expressed DsRed proteins as non-frameshift products (because there was an in-frame stop codon in the linker sequence) and DsRed-EGFP fusion proteins as frameshift products (because the stop codon was bypassed by +1 frameshifting) (Figure 3.1). Different linker sequences for +1 PRF studies in this chapter are listed in Table 3.1.

The plasmids for -1 PRF studies were created as the following. A DNA sequence GTACAAGCATCATCATCATCATTAAGA was cloned into the BsrGI/EcoRI sites in pRG3RF (Table 3.1) for the insertion of a 6X histidine tag downstream of *egfp*. This plasmid was named pRG3RFhis. Linker sequences containing -1 PRF signals (Table 3.2) were then cloned into SalI/BamHI sites in the pRG3RFhis plasmid. These linker

sequences made the downstream *egfp* in the -1 reading frame. Site direct mutagenesis was used to create control plasmids for each PRF signal according to manufacturer's protocol (Qiagen, Valencia, CA) (Table 3.2).

In the *S. cerevisiae* system, yeast strains JD932, JD1228, and JD1229 were used in the study [24]. The Kozak sequence upstream of *egfp* in the pMB2, pMBC, pO2, and pOC was first mutated by site direct mutagenesis using two primers, mut-koz and mut-kozC. These plasmids were named pMB3, pMB3C, pO3, and pO3C, respectively. The DNA fragments from *DsRed* to *egfp* in the four plasmids were then PCR amplified using two primers, pcr-MB4f and pcr-MB4r. The primer pcr-MB4f included a *SpeI* site and a Kozak sequence (TAAAC) upstream of *DsRed*. The primer pcr-MB4r included a 6X histidine tag and an *EcoRI* site downstream of *egfp*. The PCR products were digested by *SpeI* and *EcoRI* and ligated with *SpeI*/*EcoRI* digested p426GPD plasmid, a yeast shuttle vector [25], to create p426G1, p426G2, p426G3, and p426G4 (using pMB3, pMB3C, pO3, and pO3C as templates in PCR, respectively). These plasmids, p426G1, p426G2, p426G3, and p426G4, contained two *Sall* sites. To facilitate cloning of a linker flanked with *Sall* and *BamHI* sites, the *Sall* site downstream of *egfp* in p426G1 was removed by site direct mutagenesis using mut-026 and mut-026r, generating p026G1. p026G1 could be used as a backbone plasmid for replacing different PRF signals in the yeast shuttle vector. For example, yeast LA virus directed -1 PRF signal was amplified by using pcr-YLA and pcr-YLA<sub>r</sub> as PCR primers and pYDL-LA [14] as a template. The PCR product was cloned into *Sall*/*BamHI* sites in pRG026G1 to create pYDF-LA. The control plasmid for yeast LA signal (pYDF-LA0) was created by the same method expects that the PCR template is pYDL-LA0 [26].

To reverse the order of the fluorescence reporter, a different set of plasmids was

Table 3.1 Linker sequences for +1 PRF studies in Chapter 3. The P-site in the recoding site is underlined and the mutations are shown in bold.

Plasmids	Linker DNA sequence	Strain
pRG3L	TCG ACT TCT GGC TCT GGC TCT GGC GAG	RG3L (control)
pRG3RF	TCG ACT AGG GGG TAT <u>CTT</u> TGAC TAC GAG	RG3RF
pRG4NNN	TCG ACA GGG GGT <b>NNN</b> <u>CTT</u> TGAC TAC GAG	RG4NNN
pRG6GUN	TCG ACT AGG GGG <b>GUN</b> <u>CTT</u> TGAC TAC GAG	RG6GUN



Table 3.2 Linker sequences containing -1 PRF signals in Chapter 3. The P-site in the recoding site is underlined and the mutations are shown in bold.

Plasmids	Linker DNA sequence	Strain
pMB2	TCG ACT GCT AAT <u>TTT</u> TTA GGG AAG ATC TGG CCT TCC TAC AAG <u>GGA</u> AGG CCA GGG AAT TTT CTT GGA TAA AG	MB2
pMBC	TCG ACT GCT AAC <b>TT<b>C</b></b> <b>CT<b>CA</b></b> GGG AAG ATC TGG CCT TCC TAC AAG GGA AGG CCA GGG AAT TTT CTT GGA TAA AG	MBC
pO2	TCG ACT GCT AAT <u>TTT</u> TTA GGG AAG TAC TGG CCT CCG IGG GGC <u>ACG</u> AGG CCA GGC AAT TAT GTG CAG AAA CAA GTG TCC CCA TAA AG	O2
pOC	tcg act GCT AAC <b>TT<b>C</b></b> <b>CT<b>CA</b></b> GGG AAG TAC TGG CCT CCG IGG GGC ACG AGG CCA GGC AAT TAT GTG CAG AAA CAA GTG TCC CCA TAA AG	OC

constructed. The coding sequence of DsRed was PCR amplified by using Eco-mRFP and mRFP-Spe as primers and pmRFP [23] as a template. The PCR product was cloned into EcoRI/SpeI sites in pEGFP vector (Clontech) to create pgr. The BsrGI site in pgr was mutated to NheI to generate pgrL4 by using mut-pgr4 and mut-pgr4r. Yeast LA virus PRF signal was PCR amplified by using pcr-grLA and pcr-grLA<sub>r</sub> as primers and pYDF-LA or pYDF-LA0 as a template. PCR products were digested with NheI and EcoRI and ligated with NheI/EcoRI restricted pgrL4 to create pgrL4-LA (by using pYDF-LA as the template in PCR) and pgrL4-LA0 (by using pYDF-LA0 as the template in PCR). The DNA fragments from *egfp* to *DsRed* were PCR amplified by using pcr-gLA and pcr-gLA<sub>r</sub> as primers and pgrL4-LA or pgrL4-LA0 as a template. PCR products were digested with SpeI and XhoI and ligated with SpeI/XhoI restricted p026GPD to create pYGR-LA (by using pgrL4-LA as a template in PCR) and pYGR-LA0 (by using pgrL4-LA0 as a template in PCR).

### **3.4.2 Fluorescence assay**

In the *E. coli* system, cells with different plasmids were cultured in 200 µl LB medium containing 100 µg/ml ampicillin in a 96-well plate for 24 hours at 37°C and 250 rpm. Alternatively, cells can be cultured in 1 ml LB medium with the same concentration of ampicillin in a 24-well plate for 24 hours at 37°C and 250 rpm. The fluorescence was measured by a plate reader (SpectraMax M5, Molecular Devices, Sunnyvale, CA). Green fluorescence was measured with an excitation wavelength at 485nm and emission at 528nm. The red fluorescence was measured with an excitation wavelength at 530 nm and emission at 590 nm. The frameshift efficiency (FS%) was obtained as the ratio of green fluorescence to red fluorescence for test strains, normalized against the fluorescence ratio of the control strain.

The fluorescence assay for the yeast system was conducted in Dr. Dinman's laboratory at University of Maryland, College Park. The yeast cells were transformed with appropriate plasmids every time before the assay. In addition, a negative control strain (strain EV, standing for a strain with an empty vector) was transformed with p426GPD. Yeast cells were cultured in 1 ml synthetic complete medium (H-uracil) [27] in duplicate in 24 well plates at 30 °C. The fluorescence was measured by a plate reader (Synergy HT, BioTek Instruments, Inc., Winooski, VT). The green fluorescence was measured using a 485/20nm filter for the excitation wavelength and a 530/25nm filter for the emission wavelength. The red fluorescence was measured using a 512/20nm filter for the excitation wavelength and a 620/40nm filter for the emission wavelength. For the time course experiments, fluorescence signals were detected every 20 minutes for 45 hours. To account for the yeast autofluorescence [28], the fluorescence of test strains and their control strains were first subtracted with that of the EV strain (Eq.1 and Eq.2). The processed fluorescence data was then used to calculate FS% (Eq.3).

$$G_{strain}^* = \frac{G_{strain}}{OD_{strain}} - average(\frac{G_{ev}}{OD_{ev}}) \quad \text{Eq.1}$$

$$R_{strain}^* = \frac{R_{strain}}{OD_{strain}} - average(\frac{R_{ev}}{OD_{ev}}) \quad \text{Eq.2}$$

$$FS\% = \frac{average(G^* / R^*)_{test}}{average(G^* / R^*)_{control}} \times 100\% \quad \text{Eq.3}$$

where G is the green fluorescence, R is the red fluorescence, OD is the absorbance at 595 nm, and subscript strain and ev denote the interested strain and the EV strain, respectively.

### **3.5 Results**

#### **3.5.1 Using the dual reporter system in *E. coli* to study the effect of Shine Dalgarno-like sequence on RF2 frameshifting**

To study the effect of the stimulatory signal on +1 PRF by the dual fluorescence reporter system, two sets of mutants were created. In the first set, the SD-like sequence was shifted one base toward the 5' end (pRG4NNN in Table 3.1) to change the spacing between the stimulatory signals and frameshift sites. In the second set, the SD-like sequence was interrupted by changing the E-site codon in the frameshift site to GUN (pRG6GUN in Table 3.1). +1 frameshift efficiency dropped significantly when the spacing between the SD sequence and frameshift sites increased (Figure 3.2.a). +1 frameshift efficiency also dropped extensively when the SD-like sequence was interrupted (Figure 3.2.b). Therefore, the results suggest that the position, as well as the integrity, of the stimulatory signals is crucial for +1 frameshifting in RF2 expression.

#### **3.5.2 Using the dual fluorescence reporter in yeast to study the effect of anisomycin on -1 PRF**

Anisomycin was previously shown to decrease -1 PRF efficiency using a dual luciferase reporter system [24]. In the present study, the effect of anisomycin on FS% was tested in four sets of strains: (1) HIV-MB: HIV-1 group M type B (426G1) and its zero frame control (426G2); (2) HIV-O: HIV-1 group O (426G3) and its zero frame control (426G4) (3) LA-RG: yeast LA virus signal inserted into a DsRed-EGFP reporter (YDF-LA) and its zero frame control (YDF-LA0) (4) LA-GR: yeast LA virus inserted into a EGFP-DsRed reporter (YGR-LA) and its zero frame control (YGR-LA0). At stationary phase (> 20 hour culture), anisomycin decreased FS% in HIV-MB system while increased FS% in HIV-O system (Figure 3.3.a and 3.3.b). Anisomycin

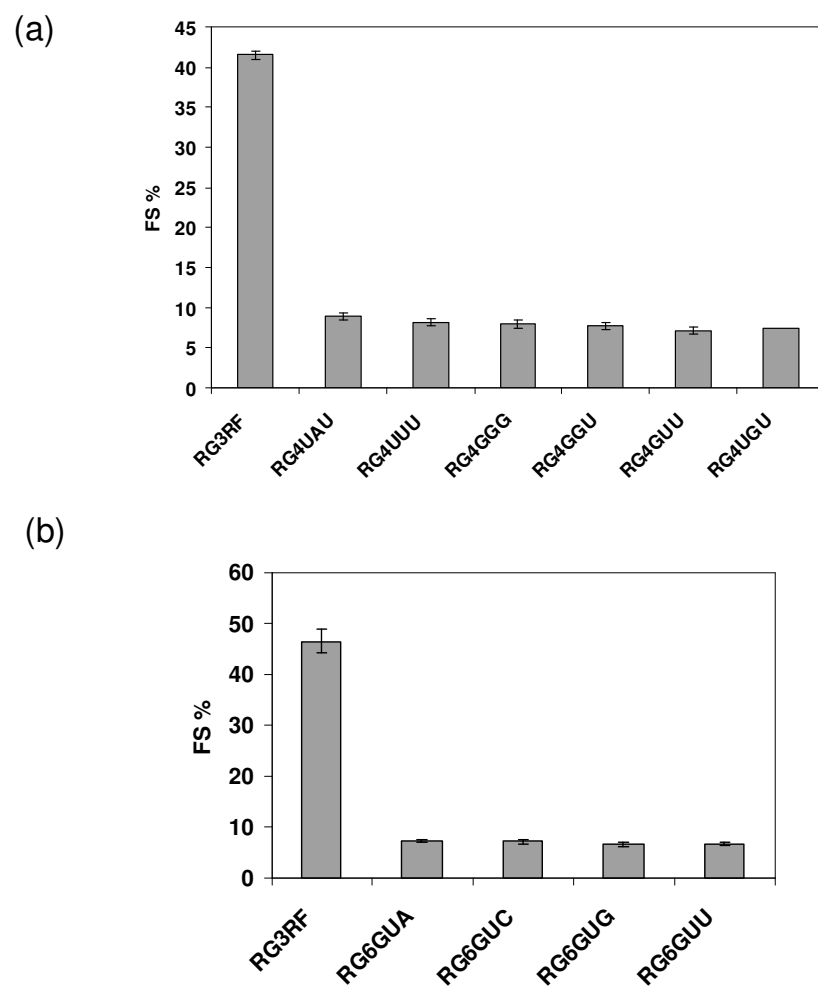


Figure 3.2. The effect of a Shine Dalgarno (SD)-like sequence on RF2 frameshifting in the *E. coli* system. Frameshift efficiency (FS%) decreases significantly when (a) the position of SD-like sequence is shifted one nucleotide upstream (b) the SD-like sequence was inserted with a guanine. Error bars represent the standard deviation. The linker sequences for these strains are listed in Table 3.1.

also decreased FS% in LA-RG and LA-GR at late stationary phase (> 35 hours) (Figure 3.3.c and 3.3.d).

### **3.5.3 Using the dual fluorescence reporter in yeast to study the effect of ribosomal protein mutations on -1 PRF**

In ribosomal protein *RPL3/TCM1*, mutations that inhibit peptidyl transferase activity were found to enhance -1 PRF efficiency in yeast cells by using a dual luciferase assay [24]. Yeast strain JD1228/JD1229 isogenic pairs in which the disrupted *RPL3/TCM1* allele is complemented with pRPL3 (wild type *RPL3/TCM1*) or pmak8-1 (Trp255Cys and Pro257Thr in *RPL3/TCM1*) [29] were used to test the effect of these mutations. At stationary phase, JD1229 resulted in lower FS% compared to JD1228 in HIV-MB system while no difference was observed in HIV-O system (Figure 3.4.a and 3.4.b). In the LA-RG system, FS% was lower in JD1229 than in JD1228 at stationary phase (Figure 3.4.c). In LA-GR system, FS% was higher in JD1229 than in JD1228 after about 22 hours (Figure 3.4.d).

## **3.6 Discussion**

### **3.6.1 Dual fluorescence in *E. coli***

The dual fluorescence reporter was successfully used in the *E. coli* system to study the effect of the SD-like sequence on RF2 frameshifting. +1 FS% decreased significantly when the spacing between the SD-like sequence and frameshift sites was increased by one nucleotide (Figure 3.2.a) or when the SD-like sequence was inserted with a guanine (Figure 3.2.b). These results are consistent with previous studies. By using  $\beta$ -galactosidase as a reporter, Weiss *et al.* observed that mutations in SD-like sequence inhibited RF2 frameshifting [30]. By using an *in vitro* translation system, Marquez *et al.* suggested that the SD:anti-SD interaction could lead to a pre-mature release of E-

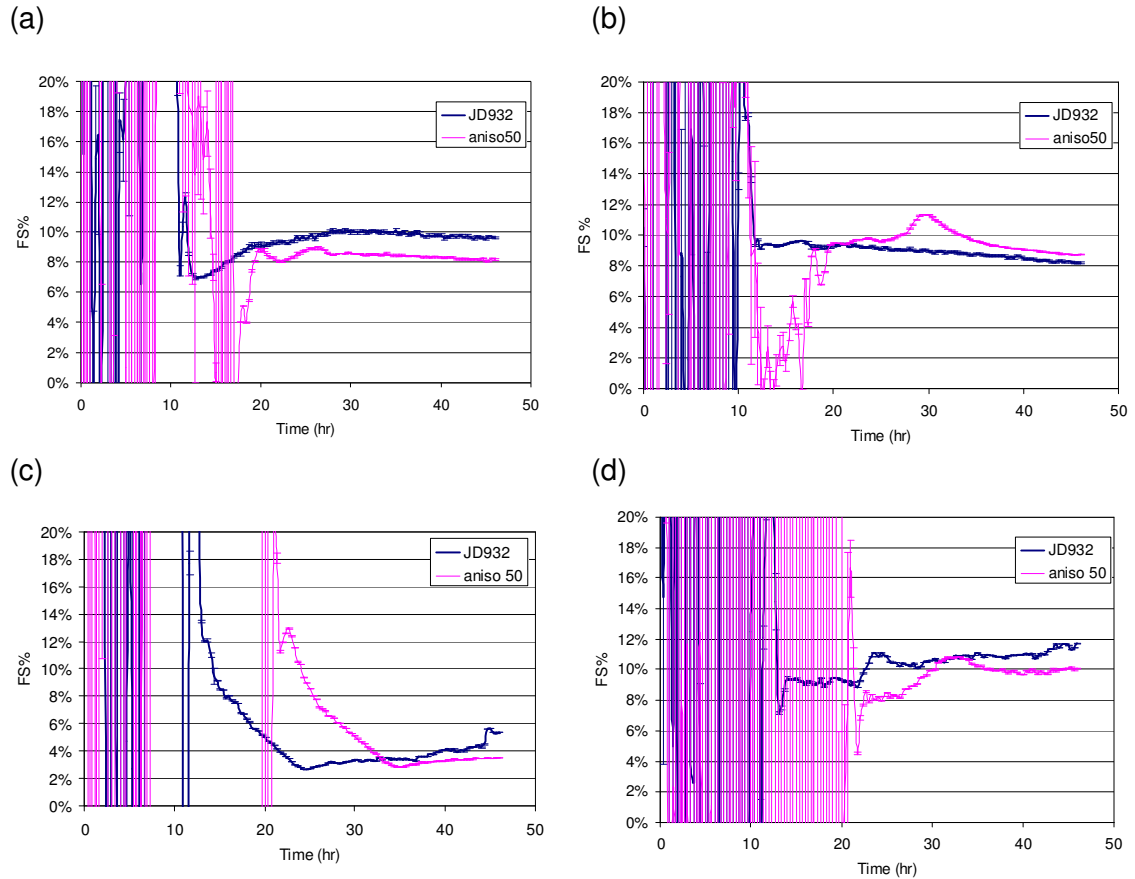


Figure 3.3. The effect of anisomycin on -1 programmed ribosomal frameshifting in four genetic backgrounds. (a) HIV-MB: HIV-1 group M type B (426G1) and its zero frame control (426G2); (2) HIV-O: HIV-1 group O (426G3) and its zero frame control (426G4) (3) LA-RG: yeast LA virus signal inserted into a DsRed-EGFP reporter (YDF-LA) and its zero frame control (YDF-LA0) (4) LA-GR: yeast LA virus inserted into a EGFP-DsRed reporter (YGR-LA) and its zero frame control (YGR-LA0). ‘JD932’ means the wild type yeast strain (JD932) cultured in the medium without anisomycin. ‘aniso 50’ means JD932 cultured in the medium with 50 µg/ml anisomycin. Error bars represent the standard deviation.

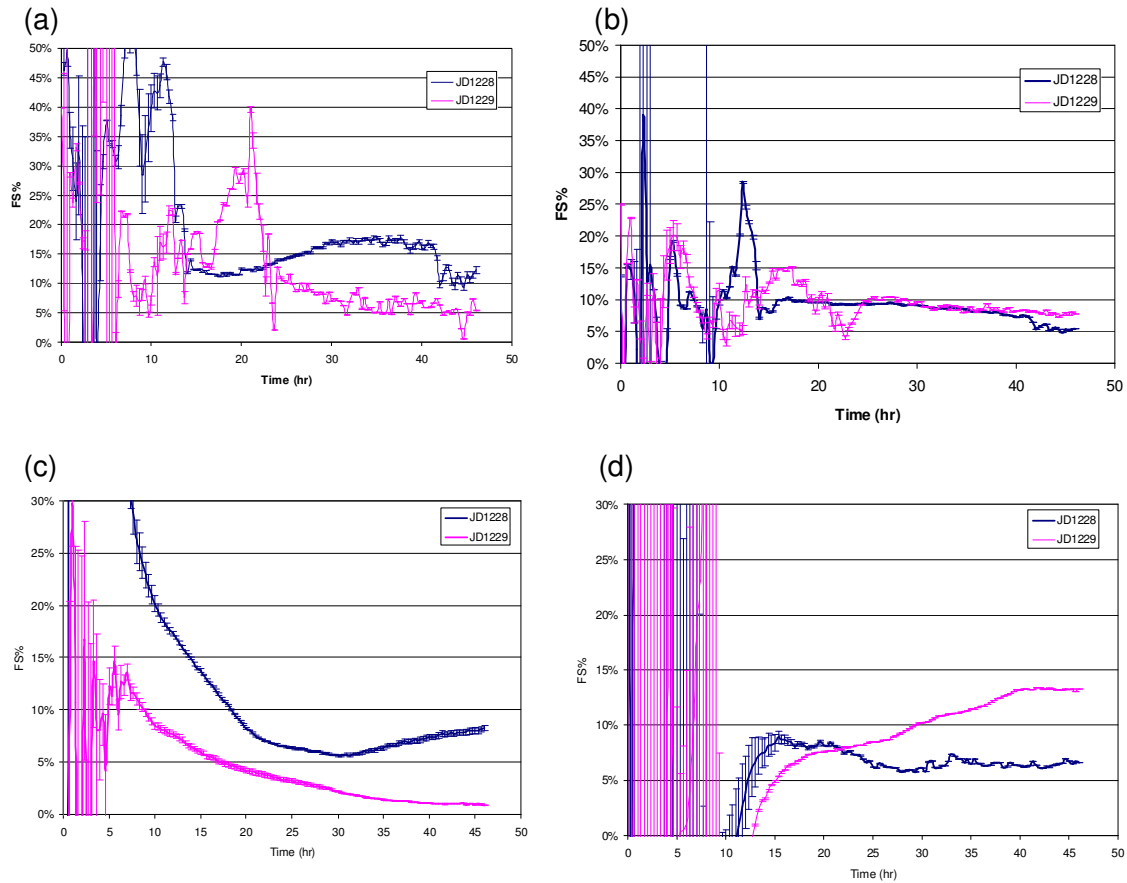


Figure 3.4. The effect of *RPL3/TCM1* mutation on  $\Delta$  programmed ribosomal frameshifting in four genetic backgrounds. (a) HIV-MB: HIV-1 group M type B (426G1) and its zero frame control (426G2); (2) HIV-O: HIV-1 group O (426G3) and its zero frame control (426G4) (3) LA-RG: yeast LA virus signal inserted into a DsRed-EGFP reporter (YDF-LA) and its zero frame control (YDF-LA0) (4) LA-GR: yeast LA virus inserted into a EGFP-DsRed reporter (YGR-LA) and its zero frame control (YGR-LA0). Yeast strain JD1228 contains wild type *RPL3/TCM1* on a plasmid and JD1229 contains mutated *RPL3/TCM1* (Trp255Cys and Pro257Thr) on a plasmid. Error bars represent the standard deviation.



site tRNA in the RF2 frameshift site [31]. When the position of the SD-like sequence was moved two or four bases upstream of the frameshift site, the E-site tRNA bound to the ribosome more stably and the incorporation of +1 frame aminoacyl-tRNA decreased dramatically [31]. Taken together, the nucleotide composition as well as the position of the SD-like sequence is critical for RF2 frameshifting in *E. coli* and the SD:antiSD interaction may exert its effort by disturbing the E-site tRNA binding, paving the way for +1 PRF.

### **3.6.2 Dual fluorescence reporter in yeast *S. cerevisiae***

Different -1 PRF signals were incorporated into the dual fluorescence reporter system in yeast and their frameshift efficiencies were monitored along the time course.

Notably, FS% fluctuated seemingly randomly and the variations between duplicates were large for the culture at times less than 20 hours (Figure 3.3 and Figure 3.4).

During the exponential growth phase, the fluorescence per OD may depend on several factors: the production of the reporter polypeptide, the maturation of the fluorescent protein, and cell division. These factors may contribute to the fluctuation in both red and green fluorescence per OD, causing an even greater variation in FS%. Therefore, frameshift efficiencies were compared during the stationary phase.

In the present work, anisomycin decreased FS% in three sets of the reporters, HIV-MB, LA-RG, and LA-GR, while increasing FS% in the HIV-O system at the stationary phase. The *RPL3/TCM1* mutations (JD1229) resulted in a lower FS% than JD1228 in the HIV-MB and LA-RG systems, no or very small change in FS% in the HIV-O setting, and a higher FS% in JD1229 than in JD1228 in LA-GR system.

Because all four systems employed -1 PRF signals, the different trends observed in the four systems were not expected. The reason why different sets of the reporter have

inconsistent results is not clear at this time. It is possible that different linker sequences alter the folding of the fusion protein differently. Because these reporter proteins need to develop the correct conformation to fluoresce, the perturbation of the protein folding may lead to an underestimation of the protein expression level.

The dual luciferase assay suggested that anisomycin decreased FS% and JD1229 containing *RPL3/TCM1* mutations enhanced FS% compared to wild type [24]. Only a subset of dual fluorescence reporter results was consistent with the dual luciferase assay (anisomycin experiments: HIV-MB, LA-RG and LA-GR; JD1228/JD1229 experiments: LA-GR). While the exact cause for the discrepancy is not known, there are several differences between a dual luciferase reporter and a dual fluorescence reporter: (1) the dual luciferase assay required collecting cells at an OD<sub>595nm</sub> of 0.7 while the dual fluorescence assay monitored FS% during the whole time course. (2) the dual luciferase assay needed to lyse cells to detect the luciferase activity and the dual fluorescence assay kept cells intact; (3) the maturation process of luciferases and fluorescent proteins are different. Firefly luciferase does not require post-translational processing for enzyme activity [32]. In contrast, fluorescent proteins mature through a multi-step process that consists of folding, chromophore formation and chromophore modification [33]. Whether these factors cause the observed discrepancy in PRF results is not clear and should be investigated further.

### ***3.7 Supplementary data***

Primers for PCR and site direct mutagenesis in this study are listed in Table 3.S1.

Table 3.S1 Primers used in Chapter 3

Name	Sequence (5' to 3')
Hind-mRFP	CCAAGCTTGATGGCCTCCTCCGAGG
mRFP-SalI	AGGTCGACGCGGCGCCGGTGGAGT
mut-koz	GGATCCCCGGGTACCGGTTGGCTATATGGTGAGCAAGGGCGAGG
mut-kozc	CCTCGCCCTTGCTCACCATATAGCCAACCGGTACCCGGGGATCC
pcr-MB4f	TTACTAGTTAAACATGGCCTCCTCCGAGGACGTCATC
pcr-MB4r	AAGAATTCTTAATGATGATGATGATGATGCTTGTACAGC
mut-026	CAAGCTTATCGATACCGTCTACCTCGAGTCATGTAATTAG
mut-026r	CTAATTACATGACTCGAGGTAGACGGTATCGATAAGCTTG
pcr-YLA	TTTGTCGACCACTTCTAGGATCAATGCG
pcr-YLA <sub>r</sub>	TTTGGATCCAAAATTAAGGGATCGGTACCCCCGGG
Eco-mRFP	CCCAGAATTCATGGCCTCCTCCGAGG
mRFP-Spe	TTTACTAGTTTAGGCGCCGGTGGAG
mut-pgr4	CTCGGCATGGACGAGCTAGCAGTCTGGCTCTGGCTCTGGC
mut-pgr4 <sub>r</sub>	GCCAGAGCCAGAGCCAGACTGCTAGCTCGTCCATGCCGAG
pcr-grLA	CGAGCTAGCTACTTCTAGGATCAATGCG
pcr-grLA <sub>r</sub>	CCATGAATTCGGGATCGGTACCCC
pcr-gLA	TTACTAGTTAAACATGGTGAGCAAGGGCGAGG
pcr-gLA <sub>r</sub>	AACTCGAGTTAATGATGATGATGATGATGGGCGCCGGTGGAGTGGCGGCCC

### 3.8 Conclusion

A dual fluorescence reporter system was developed in *E. coli* and *S. cerevisiae*.

Different PRF sites were inserted between two fluorescence reporter genes, monomeric DsRed and EGFP. For +1 PRF studies in *E. coli*, the results suggested that the position as well as the integrity of the SD-like sequence in RF2 frameshift site were crucial for +1 PRF in *E. coli*. For -1 PRF studies in yeast, the frameshift efficiency under different genetic backgrounds and culture conditions was monitored in real time. However, the dual fluorescence assay data were not completely consistent with the dual luciferase reporter assay results. Therefore, the yeast dual fluorescence reporter system still needs to be improved further.

## REFERENCES

1. Kurland, C.G. (1992) Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.*, **26**, 29-50.
2. Sundararajan, A., Michaud, W.A., Qian, Q., Stahl, G. and Farabaugh, P.J. (1999) Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. *Mol. Cell*, **4**, 1005-1015.
3. Farabaugh, P.J. (1996) Programmed translational frameshifting. *Annu. Rev. Genet.*, **30**, 507-528.
4. Gesteland, R.F. and Atkins, J.F. (1996) Recoding: Dynamic reprogramming of translation. *Annu. Rev. Biochem.*, **65**, 741-768.
5. Jacks, T., Power, M.D., Masiarz, F.R., Luciw, P.A., Barr, P.J. and Varmus, H.E. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280-283.
6. Baranov, P.V., Henderson, C.M., Anderson, C.B., Gesteland, R.F., Atkins, J.F. and Howard, M.T. (2005) Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology*, **332**, 498-510.
7. Dinman, J.D., Ruiz-Echevarria, M.J. and Peltz, S.W. (1998) Translating old drugs into new treatments: Ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.*, **16**, 190-196.
8. Belcourt, M.F. and Farabaugh, P.J. (1990) Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell*, **62**, 339-352.
9. Farabaugh, P.J., Zhao, H. and Vimaladithan, A. (1993) A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell*, **74**, 93-103.
10. Curran, J.F. (1993) Analysis of effects of tRNA:Message stability on frameshift frequency at the *Escherichia coli* RF2 programmed frameshift site. *Nucleic Acids Res.*, **21**, 1837-1843.
11. Baranov, P.V., Gesteland, R.F. and Atkins, J.F. (2002) Release factor 2 frameshifting sites in different bacteria. *EMBO Rep.*, **3**, 373-377.

12. Hansen, T.M., Baranov, P.V., Ivanov, I.P., Gesteland, R.F. and Atkins, J.F. (2003) Maintenance of the correct open reading frame by the ribosome. *EMBO Rep.*, **4**, 499-504.
13. Grentzmann, G., Ingram, J.A., Kelly, P.J., Gesteland, R.F. and Atkins, J.F. (1998) A dual-luciferase reporter system for studying recoding signals. *RNA*, **4**, 479-486.
14. Harger, J.W. and Dinman, J.D. (2003) An *in vivo* dual-luciferase assay system for studying translational recoding in the yeast *Saccharomyces cerevisiae*. *RNA*, **9**, 1019-1024.
15. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. and Prasher, D.C. (1994) Green fluorescent protein as a marker for gene expression. *Science*, **263**, 802-805.
16. Tsien, R.Y. (1998) The green fluorescent protein. *Annu. Rev. Biochem.*, **67**, 509-544.
17. Cormack, B.P., Valdivia, R.H. and Falkow, S. (1996) FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, **173**, 33-38.
18. Matz, M.V., Fradkov, A.F., Labas, Y.A., Savitsky, A.P., Zaraisky, A.G., Markelov, M.L. and Lukyanov, S.A. (1999) Fluorescent proteins from nonbioluminescent anthozoa species. *Nat. Biotechnol.*, **17**, 969-973.
19. Gurskaya, N.G., Fradkov, A.F., Terskikh, A., Matz, M.V., Labas, Y.A., Martynov, V.I., Yanushevich, Y.G., Lukyanov, K.A. and Lukyanov, S.A. (2001) GFP-like chromoproteins as a source of far-red fluorescent proteins. *FEBS Lett.*, **507**, 16-20.
20. Shaner, N.C., Steinbach, P.A. and Tsien, R.Y. (2005) A guide to choosing fluorescent proteins. *Nat. Methods*, **2**, 905-909.
21. Choe, J., Guo, H.H. and van den Engh, G. (2005) A dual-fluorescence reporter system for high-throughput clone characterization and selection by cell sorting. *Nucleic Acids Res.*, **33**, e49.
22. Pflieger, B.F., Pitera, D.J., Smolke, C.D. and Keasling, J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027-1032.
23. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 7877-7882.
24. Plant, E.P., Perez-Alvarado, G.C., Jacobs, J.L., Mukhopadhyay, B., Hennig, M. and Dinman, J.D. (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, e172

25. Mumberg,D., Muller,R. and Funk,M. (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene*, **156**, 119-122.
26. Muldoon-Jacobs,K.L. and Dinman,J.D. (2006) Specific effects of ribosome-tethered molecular chaperones on programmed -1 ribosomal frameshifting. *Eukaryot. Cell.*, **5**, 762-770.
27. Dinman,J.D. and Wickner,R.B. (1994) Translational maintenance of frame: Mutants of *Saccharomyces cerevisiae* with altered -1 ribosomal frameshifting efficiencies. *Genetics*, **136**, 75-86.
28. Knight,A.W., Goddard,N.J., Fielden,P.R., Barker,M.G., Billinton,N. and Walmsley,R.M. (1999) Fluorescence polarization of green fluorescent protein (GFP). A strategy for improved wavelength discrimination for GFP determinations. *Anal. Commun.*, **36**, 113-117.
29. Meskauskas,A., Harger,J.W., Jacobs,K.L. and Dinman,J.D. (2003) Decreased peptidyltransferase activity correlates with increased programmed -1 ribosomal frameshifting and viral maintenance defects in the yeast *Saccharomyces cerevisiae*. *RNA*, **9**, 982-992.
30. Weiss,R.B., Dunn,D.M., Dahlberg,A.E., Atkins,J.F. and Gesteland,R.F. (1988) Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.*, **7**, 1503-1507.
31. Marquez,V., Wilson,D.N., Tate,W.P., Triana-Alonso,F. and Nierhaus,K.H. (2004) Maintaining the ribosomal reading frame: The influence of the E site during translational regulation of release factor 2. *Cell*, **118**, 45-55.
32. Kolb,V.A., Makeyev,E.V. and Spirin,A.S. (1994) Folding of firefly luciferase during translation in a cell-free system. *EMBO J.*, **13**, 3631-3637.
33. Miyawaki,A., Nagai,T. and Mizuno,H. (2003) Mechanisms of protein fluorophore formation and engineering. *Curr. Opin. Chem. Biol.*, **7**, 557-562

## CHAPTER 4

### A NEW KINETIC MODEL REVEALS THE SYNERGISTIC EFFECT OF E-, P-, AND A- SITES ON +1 RIBOSOMAL FRAMESHIFTING

#### **4.1 Preface**

This chapter is adapted from Liao, P.Y., Gupta, P., Petrov, A., Dinman, J.D., Lee, K.H. 2008. A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Research*. 36: 2619-2629. This study describes a kinetic model for +1 programmed ribosomal frameshifting. The model predictions are tested experimentally using a dual fluorescence reporter system in *Escherichia coli*. The results suggest that ribosome E-, P-, and A-sites are all involved in +1 frameshifting.

#### **4.2 Abstract**

Programmed ribosomal frameshifting (PRF) is a process by which ribosomes produce two different polypeptides from the same mRNA. In this study, we propose three different kinetic models of +1 PRF, incorporating the effects of the ribosomal E-, P-, and A-sites toward promoting efficient +1 frameshifting in *Escherichia coli*. Specifically, the timing of E-site tRNA dissociation is discussed within the context of the kinetic proofreading mechanism of aminoacylated tRNA (aa-tRNA) selection. Mathematical modeling using previously determined kinetic rate constants reveals that destabilization of deacylated tRNA in the E-site, rearrangement of peptidyl-tRNA in the P-site, and availability of cognate aa-tRNA corresponding to the A-site act synergistically to promote efficient +1 PRF. The effect of E-site codon:anticodon interactions on +1 PRF was also experimentally examined with a dual fluorescence reporter construct. The combination of predictive modeling and empirical testing

allowed the rate constant for P-site tRNA slippage ( $k_s$ ) to be estimated as  $k_s \approx 1.9 \text{ s}^{-1}$  for release factor 2 (RF2) frameshifting sequence. These analyses suggest that P-site tRNA slippage is the driving force for +1 ribosomal frameshifting while the presence of a “hungry codon” in the A-site and destabilization in the E-site further enhance +1 PRF in *E. coli*.

### 4.3 Introduction

Programmed ribosomal frameshifting (PRF) is a coded shift in reading frame during translation of an mRNA transcript. Consequently, one transcript may yield two different protein products, an inframe product and a frameshift product. PRF has been observed to occur in various organisms including prokaryotes and eukaryotes. In +1 PRF, the ribosome skips over one nucleotide toward 3' direction. +1 PRF has been observed in *Escherichia coli* in the translation of *prfB* to produce release factor 2 (RF2) [1]. In *Saccharomyces cerevisiae* two retrotransposable elements, *Ty1* and *Ty3* [2,3], and three genes, *ABP140* [4], *EST3* [5], and *OAZ1* [6] use +1 PRF. The expression of mammalian antizyme has also been shown to involve +1 PRF [7].

Several features have been shown to facilitate +1 PRF: (1) low levels of aminoacylated-tRNA (aa-tRNA) corresponding to the in-frame A-site codon, *i.e.* hungry codons [8]; (2) the ability of P-site tRNA to form near-cognate interactions with the shifted frame codon, *i.e.* slippery sequence [9]; and (3) the presence of a stimulatory signal, such as a Shine-Dalgarno (SD)-like sequence upstream of the frameshift site [10] or a RNA secondary structure downstream of the frameshift site [3]. Both (1) and (3) may promote a pause in translation elongation, which allows more time for a recoding event to occur, suggesting that +1 PRF is kinetically driven [11].



Several mechanistic models have been proposed to explain +1 PRF [11-13]. The kinetic model of Baranov *et al.* [13] illustrated the dependence of frameshift efficiency on the stability of the P-site interaction and the concentration of incoming aa-tRNA available for the zero and +1 frames. This kinetic model is consistent with observations from several frameshifting studies. For example, the codon: anticodon interaction in the +1 frame of the P-site has been shown to affect the amount of frameshift products [9]. Overexpression of the cognate P-site tRNAs has also been shown to dramatically reduce +1 PRF in yeast and vice versa [2,14,15].

Recent experimental observations suggest that the E-site plays a crucial role in the efficiency of +1 PRF in *E. coli* [16]. In that study, premature release of E-site tRNA from the ribosome correlated with high levels of frameshifting products. A mutagenesis study of 23S rRNA has also illustrated the correlation between E-site tRNA binding and the maintenance of reading frame [17]. A recently published study shows that RF2 programmed frameshifting is inversely correlated with the E-site stability in *E. coli* [18]. To date, no published kinetic model of +1 PRF has explained the effect of E-site tRNA release on +1 PRF.

In the present study, we propose a new mathematical model for +1 PRF in *E. coli*, which incorporates the effects of E-, P-, and A-site interactions in promoting high levels of frameshifting. Previously published theories of +1 PRF usually focus on a single aspect of +1 PRF (*e.g.* A-site tRNA abundance, stability of P-site tRNA - ribosome interaction and etc. [8,9,14,15]). Here, we present a model synthesizing previously observed effects of all three ribosomal tRNA binding sites on +1 PRF efficiency in *E. coli*. Of particular note, this is the first model combining the concepts of kinetic proofreading of aa-tRNA selection [19] with the allosteric model [20] to

describe +1 PRF. The proposed mathematical model suggests that the rate of P-site tRNA slippage is the most significant parameter in the +1 PRF event, while the abundance of cognate aa-tRNA and the rate of E-site tRNA release act synergistically to promote highly efficient +1 PRF.

#### ***4.4 Kinetic model***

An elegant series of biochemical studies have contributed to a very detailed kinetic model of A-site tRNA selection [19]. In this model, fast initial binding of the ternary complex EF-Tu:aa-tRNA:GTP is followed by codon recognition. Codon recognition triggers EF-Tu GTPase activation, which leads to the GTP hydrolysis and dissociation of EF-Tu from the ribosome. Factor dissociation is followed by the spontaneous accommodation of the acceptor end of the aa-tRNA into the A-site or the rejection of the aa-tRNA by proofreading. This concept is illustrated along the top of Figure 4.1.

Other recent studies suggested that events at the ribosomal E-site are involved in coordinating this process, specifically that E-site tRNA dissociation occurs prior to GTP hydrolysis [21]. Functional studies suggest that +1 PRF efficiency is linked to the E-site occupation and the identity of the E-site tRNA [16-18]. Following the allosteric model of the elongation cycle, the E-site is occupied at the start of each cycle prior to aa-tRNA selection, and A-site tRNA binding promotes release of the E-site tRNA, followed in turn by peptidyltransfer and translocation. During translocation the deacylated tRNA is shifted from the P-site to the E-site. Thus there are two events that affect the E-site occupancy: aa-tRNA selection and translocation. The previously observed effects of tRNA abundance and amino acid starvation on +1 PRF efficiency strongly suggest that +1 PRF occurs during A-site tRNA selection [2]. Importantly, recent X-ray crystal structures show that the E-site tRNA can form 1-3 base pairing

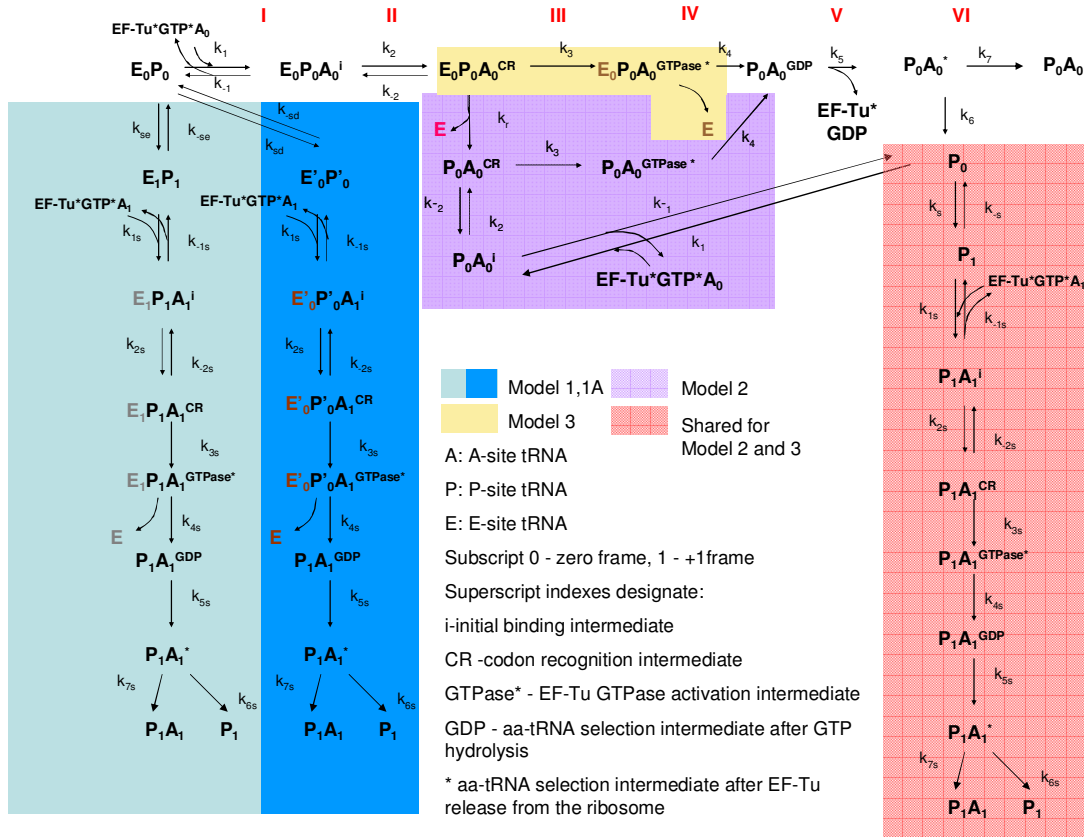


Figure 4.1. The three kinetic models for +1 PRF in *E. coli*. Steps I-VI illustrate the non-frameshifting translation elongation process: I. initial binding; II. codon recognition; III. GTPase activation; IV. GTP hydrolysis; V. EF-Tu dissociation; VI. accommodation. In Model 1, both E-site and P-site tRNAs slip into the +1 frame and follow the +1 frame aa-tRNA selection to produce frameshift proteins. ( $P_1A_1$ ). In Model 1A, both E-site and P-site tRNAs are destabilized by stimulatory signals and follow the +1 frame aa-tRNA selection to produce frameshift proteins. ( $P_1A_1$ ). Model 2 and 3 differ in the timing of the E-site tRNA dissociation step (In Model 2, E-site tRNA dissociation occurs during the codon recognition step while in Model 3, E-site tRNA dissociates after codon recognition). Both Models (2 and 3) result in the formation of ribosomes with only P-site tRNA ( $P_0$ ), which can slip to the +1 frame to form  $P_1$  and result in the formation of frameshift proteins ( $P_1A_1$ ).

interactions with the mRNA [22,23]. Thus E-site tRNA destabilization may make ribosomes more prone to frameshifting by reducing the extent of tRNA:mRNA interactions.

Because the exact timing of dissociation is unknown, three different models of +1 PRF in *E. coli* that differ in the timing of E-site tRNA release (Figure 4.1) were constructed in the present study. In Model 1, simultaneous slippage of E- and P-site tRNAs is hypothesized to occur before aa-tRNA selection (shaded in blue). In this model, the rate constant of the simultaneous slippage ( $k_{se}$ ) is determined by the stability of the  $E_0P_0$  complex (ribosomes with E- and P-site occupied). The stability of this complex depends on the identity of the P-site tRNA [13,14], and to some extent on the E-site tRNA [24,25]. The 3' slippage results in the formation of  $E_1P_1$  complex, in which both E-site and P-site tRNAs have been shifted by one base. Thereafter, the ribosome can follow the normal elongation cycle to produce frameshift proteins ( $P_1A_1$ ). Similarly, in Model 1A (shaded in dark blue in Figure 4.1), stimulatory signals may destabilize  $E_0P_0$ , yielding an unstable complex  $E'_0P'_0$ . As this destabilization occurs, the codon at the A site is shifted, leaving both zero frame aa-tRNA ( $A_0$ ) and +1 frame aa-tRNA ( $A_1$ ) as near-cognate ternary complexes. The binding of  $A_1$  to the A site will trigger the release of the E-site tRNA. This step is then followed by the slippage of the P-site tRNA to base pair with the +1 frame. Frameshift products ( $P_1A_1$ ) would then be produced by following the remaining steps of aa-tRNA selection by ribosomes.

Slippage could also occur during aa-tRNA selection. To accommodate for the unclear timing of E-site tRNA release, two additional models are proposed. In Model 2, E-site tRNA dissociation occurs during the codon recognition step. E-site empty ribosomes formed at this step can either continue with the subsequent steps of aa-tRNA selection

or undergo the reverse reaction to yield initial binding complex  $P_0A_0^i$ .  $P_0A_0^i$  can again undergo the aa-tRNA selection or release the aa-tRNA to form ribosomes with only P-site tRNA occupied ( $P_0$ ). Depending upon the slippage constant ( $k_s$ ), tRNA in the  $P_0$  state can slip to base pair with the +1 frame and form the  $P_1$  state.  $P_1$  can then go through the +1 frame aa-tRNA selection and produce the frameshift proteins ( $P_1A_1$ ). Alternatively, the E-site tRNA might dissociate after codon recognition (Model 3). In this model, E-site empty ribosomes ( $P_0$ ) can be formed consequent to aa-tRNA rejection during the accommodation step. Importantly, because the initial binding of aa-tRNA is fast and nonspecific, Model 2 would result in the formation of a significantly larger fraction of the ribosomes in  $P_0$  states as compared to Model 3.

## **4.5 Materials and methods**

### **4.5.1 Computation of the kinetic model**

All three models were mathematically described by systems of ordinary differential equations (see Text in Supplementary Data). Assuming steady state, the expressions of intermediate concentrations in terms of initial reactant ( $E_0P_0$ ) were solved by Matlab 7.2 (Mathworks Inc., USA). By applying the empirically-determined rate constants and assumed ranges of rate constants of P-site tRNA slippage, and rate constants of E-site tRNA release (Table 4.S1 and 4.S2 in Supplementary Data) with different aa-tRNA concentrations (Table 4.S3 in Supplementary Data), the amount of non-frameshift proteins ( $P_0A_0$ ) and frameshift proteins ( $P_1A_1$ ) were calculated. The frameshift efficiency (FS%) in the model is defined as the ratio of  $P_1A_1$  to total proteins ( $P_0A_0 + P_1A_1$ ) multiplied by 100 %.

### **4.5.2 Plasmids and bacterial strains**

*Escherichia coli* XL1 blue MRF' (Stratagene) was used in all experimental studies.

The gene sequence of monomeric DsRed [26] was first cloned between HindIII and Sall sites in pEGFP vector (Clontech, Mountain View, CA) to create pRG plasmid, which can express DsRed-EGFP fusion protein. Different linker sequences were made from complementary oligonucleotides (Integrated DNA Technology, Coralville, IA, USA) and were cloned between Sall and BamHI sites between the coding sequence of DsRed and EGFP in the pRG plasmid. The linker sequence for the control strain is tcgacttctggctctggctctggcgag, which kept both DsRed and EGFP coding sequences in frame. The linker sequences for the mutants contained mutated RF2 frameshift sites (tcgactagggggUNNctttgactacgag) which made EGFP coding sequence in +1 frame (UNN refers to the E-site codon when +1 frameshifting is taking place and the stop codon is underlined). The control strain expressed only the DsRed-EGFP fusion protein. The mutants expressed DsRed proteins as non-frameshift proteins (because of the stop codon in the linker sequence) and DsRed-EGFP fusion protein as frameshift proteins (because the stop codon is bypassed by +1 frameshifting). Thirteen mutants differing only in the E-site codon (UNN) in the recoding sites were constructed. Among the thirteen mutants, the first base in the E-site codon was kept intact to maintain SD-like sequence and stop codons were avoided.

#### **4.5.3 Fluorescence assay**

Cells with different plasmids were cultured in 200 µl Luria-Bertani (LB) medium containing 100 µg/ml ampicillin in a 96-well plate for 24 hours at 37°C, 250 rpm. The fluorescence was then measured by plate reader (SpectraMax Gemini EM, Molecular Devices). The green fluorescence was measured with excitation wavelength at 485 nm and emission at 528 nm. The red fluorescence was measured with excitation wavelength at 530 nm and emission at 590 nm. From the fluorescence measurement, the experimental frameshift efficiency ( $FS\%_{exp}$ ) was obtained as the ratio of green

fluorescence to red fluorescence for the mutant strains (containing RF2 sequence with different E-site codons), normalized against the fluorescence ratio of the control strain.

#### 4.5.4 Chi-square analysis

Chi-square is defined as:  $\chi^2 = \sum_i \left( \frac{FS\%^i - FS\%_{\text{exp}}^i}{SD_{\text{exp}}^i} \right)^2$ , where  $i$  refers to different E-site codons ( $i = 1-13$ , for 13 tested E-site codons),  $FS\%$  is the frameshift efficiency calculated by the model and  $FS\%_{\text{exp}}$  is the frameshift efficiency observed in the experiment. The rate constant of E-site tRNA release,  $k_r$ , was assumed as  $k_r = A' \exp(-m_j \Delta G_c / RT)$ , where  $A'$  is the pre-exponential constant for the effect of the stimulatory signals (the same for all tested E-site codons);  $\Delta G_c$  is the codon:anticodon interaction in the E-site [27];  $m_j$  is the modifying factor to account for other factors (*e.g.* tRNA:ribosome interactions, base modification etc.) that may affect the contribution of the base pairing on  $k_r$  ( $j=1-6$ , see Table 4.S4 in Supplementary Data);  $R$  is the gas constant ( $8.314 \text{ J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$ );  $T$  is the temperature (310K). Matlab V.7.2 was used to optimize the values of  $k_s$ ,  $A'$  and  $m_j$  that resulted in the minimum chi-square value.

### 4.6 Results

#### 4.6.1 Mathematical model.

The three major variables in the model are the rate constant of P-site tRNA slippage ( $k_s$ ), the rate constant of E-site tRNA release ( $k_r$ ), and the concentration of cognate aa-tRNA for zero-frame codon in the A-site ( $\text{cog.A}_0$ ). To understand the synergistic effect of  $k_s$ ,  $k_r$  and  $\text{cog.A}_0$ , surface plots are used to show the effect of any two parameters on  $FS\%$  while keeping the third parameter as a constant. Figure 4.2.a shows the effect of  $k_s$  and  $k_r$  on  $FS\%$ . An increase in  $FS\%$  is observed as  $k_r$  and  $k_s$  are increased. Figure 4.2.b shows an example of the synergistic effect of E-site ( $k_r$ ) and P-

site ( $k_s$ ): while a 10 fold increase in  $k_r$  or  $k_s$  alone results in an increase in FS%, a 10 fold increase in both parameters results in a greater increase in FS% than the summation of the individual effects. Figure 4.2.c and Figure 4.2.d show the cross section curves of Figure 4.2.a. This analysis suggests that the effect of  $k_r$  is more significant when  $k_r$  is below  $10 \text{ s}^{-1}$  (Figure 4.2.c). Interestingly, the effect of  $k_r$  on FS% is less important for smaller values of  $k_s$  (only 1 % increase in FS% with increasing  $k_r$  for  $k_s = 0.05 \text{ s}^{-1}$ ), which suggests that the effect of the E-site tRNA release becomes prominent above a threshold value of P-site tRNA slippage (represented by  $k_s$ ).

Additionally, the model reveals a synergistic effect of P-site tRNA slippage and the hungry codon (Figure 4.3.a). The analysis suggests that FS% increases as  $\text{cog.A}_0$  decreases and as  $k_s$  increases. Figure 4.3.b shows an example of the synergistic effect of P-site ( $k_s$ ) and A-site ( $\text{cog.A}_0$ ): while a 10 fold decrease in  $\text{cog.A}_0$  or a 10 fold increase in  $k_s$  results in an increase in FS%, a 10 fold change in both parameters results in a greater increase in FS% than the summation of the individual effects. Figure 4.3.c and Figure 4.3.d show the cross section curves of Figure 4.3.a. Importantly, the effect of  $\text{cog.A}_0$  on FS% decreases with  $k_s$  (Figure 4.3.c). As a result, the hungry codon effect (represented by a small value of  $\text{cog.A}_0$ ) becomes more significant as the probability of P-site tRNA slippage increases (represented by larger  $k_s$ ).

The model also shows the synergistic effect between hungry codon at the A-site and release of tRNAs from the E-site. Examination of Figure 4.2.c and Figure 4.3.c shows that the effects of  $k_r$  and  $\text{cog.A}_0$  become significant only for higher values of  $k_s$ . Therefore, a higher value of  $k_s$  ( $5 \text{ s}^{-1}$ ) was chosen to study the effect of  $\text{cog.A}_0$  and  $k_r$  on FS% (Figure 4.4.a). The analysis shows that FS% increases as  $k_r$  increases and as



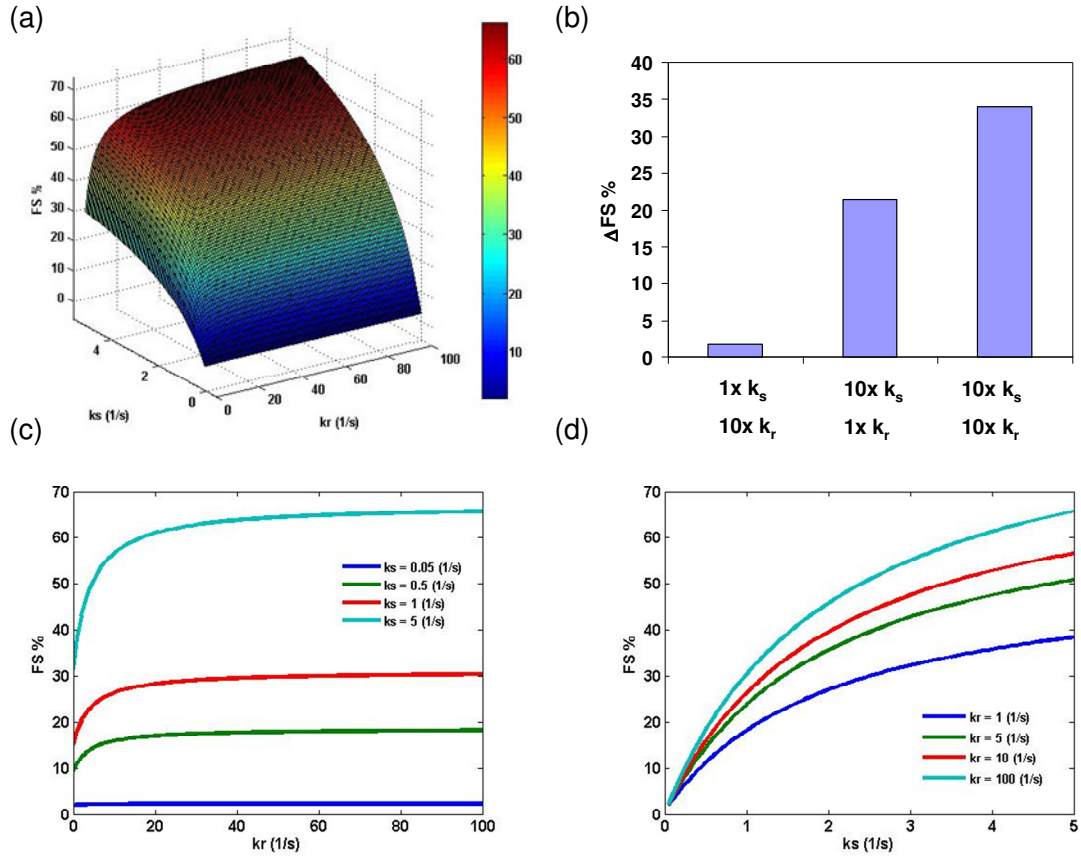


Figure 4.2. (a) The effect of P-site tRNA slippage (represented by  $k_s$ ) and E-site tRNA release (represented by  $k_r$ ) on FS% at fixed concentration of zero-frame cognate aa-tRNA ( $\text{cog.A}_0 = 1\%$ ). All the other parameters are assumed to be constants (Table 4.S1. in Supplementary Data). (b) An example of the synergistic effect of E-site ( $k_r$ ) and P-site ( $k_s$ ) (data points from Figure 4.2.a). 1x means the parameter is the same as a randomly chosen base point ( $k_s = 0.2 \text{ s}^{-1}$ ,  $k_r = 1 \text{ s}^{-1}$ ). 10x means a 10 fold increase in the parameter.  $\Delta$ FS% refers to the increase in FS% as compared to the base point. (c) The effect of  $k_r$  on FS% at different values of  $k_s$  ( $0.05$ - $5 \text{ s}^{-1}$ ). (d). The effect of  $k_s$  on FS% at different values of  $k_r$  ( $1$ - $100 \text{ s}^{-1}$ ).

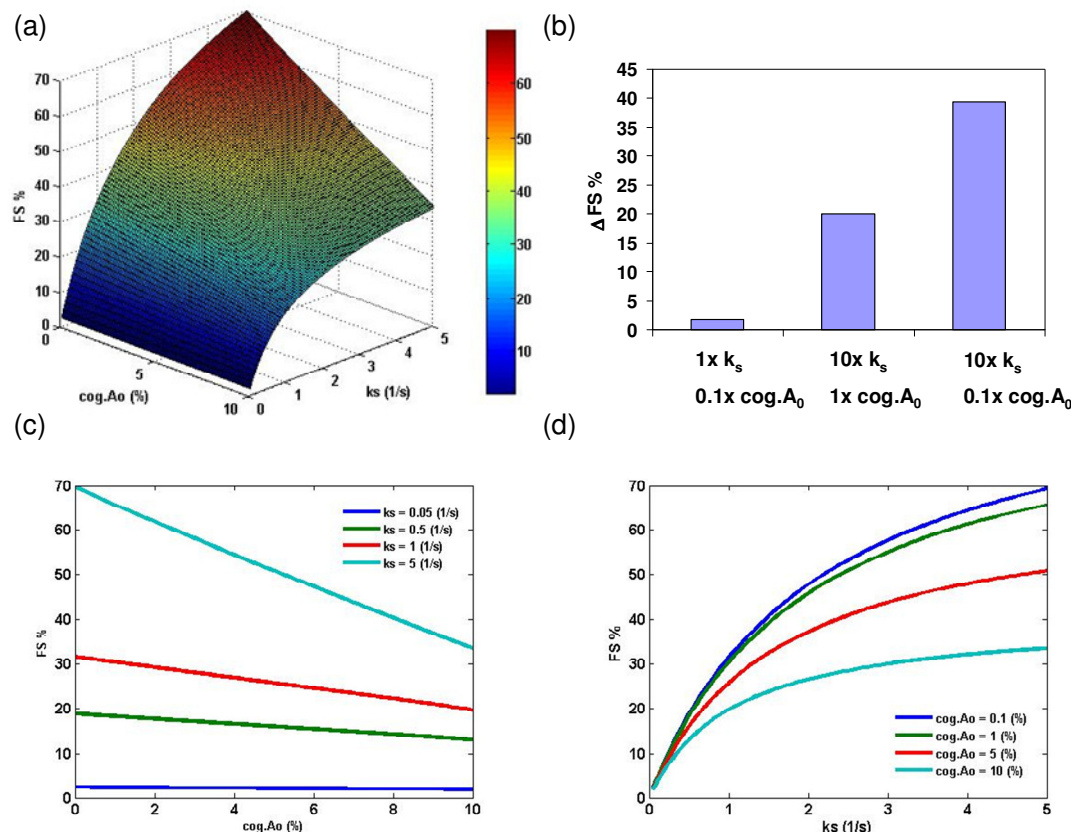


Figure 4.3. (a) The effect of P-site tRNA slippage (represented by  $k_s$ ) and the concentration of zero-frame cognate aa-tRNA ( $\text{cog.A}_0$ ) on FS% at fixed rate constant of E-site tRNA release ( $k_r = 100 \text{ s}^{-1}$ ). All the other parameters are assumed to be constants (Table 4.S1. in Supplementary Data). (b) An example of the synergistic effect of P-site ( $k_s$ ) and A-site ( $\text{cog.A}_0$ ) (data points from Figure 4.3.a). 1x means the parameter was the same as a randomly chosen base point ( $k_s = 0.2 \text{ s}^{-1}$ ,  $\text{cog.A}_0 = 10 \%$ ). 10x means a 10 fold increase in the parameter. 0.1x means a 10 fold decrease in the parameter.  $\Delta\text{FS}\%$  refers to the increase in FS% as compared to the base point. (c) The effect of  $\text{cog.A}_0$  on FS% at different values of  $k_s$  (0.05-5  $\text{s}^{-1}$ ). (d) The effect of  $k_s$  on FS% at different values of  $\text{cog.A}_0$  (0.1-10%).

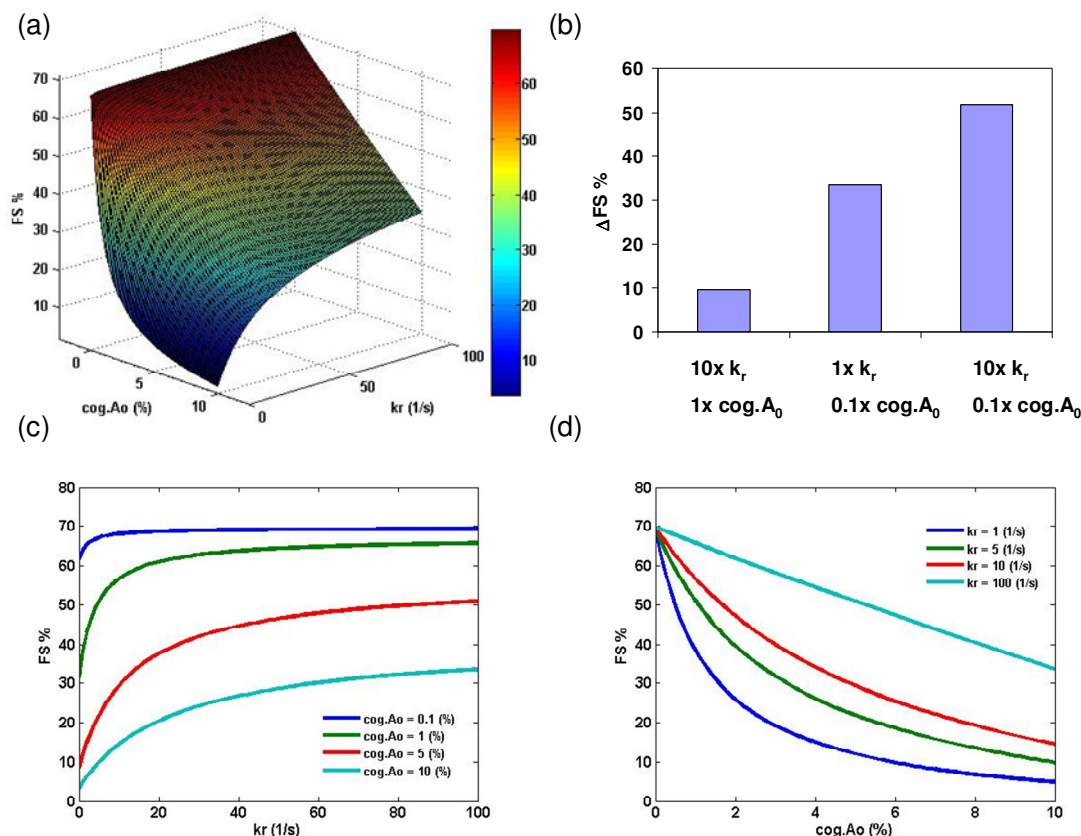


Figure 4.4. (a) The effect of the concentration of zero-frame cognate aa-tRNA ( $\text{cog.A}_0$ ) and E-site tRNA release (represented by  $k_r$ ) on FS% at fixed rate constant of P-site tRNA slippage ( $k_s = 5 \text{ s}^{-1}$ ). All the other parameters are assumed to be constants (Table 4.S1. in Supplementary Data). (b) An example of the synergistic effect of E-site ( $k_r$ ) and A-site ( $\text{cog.A}_0$ ) (data points from Figure 4.4.a). 1x means the parameter was the same as a randomly chosen base point ( $k_r = 1 \text{ s}^{-1}$ ,  $\text{cog.A}_0 = 10 \%$ ). 10x means a 10 fold increase in the parameter. 0.1x means a 10 fold decrease in the parameter.  $\Delta\text{FS}\%$  refers to the increase in FS% as compared to the base point. (c) The effect of  $k_r$  on FS% at different values of  $\text{cog.A}_0$  (0.1-10%). (d) The effect of  $\text{cog.A}_0$  on FS% at different values of  $k_r$  (1-100  $\text{s}^{-1}$ ).

cog.A<sub>0</sub> decreases, respectively. Figure 4.4.b shows an example of the synergistic effect of E-site ( $k_r$ ) and A-site (cog.A<sub>0</sub>): while a 10 fold increase in  $k_r$  or a 10 fold decrease in cog.A<sub>0</sub> results in an increase in FS%, a 10 fold change in both parameters results in a greater increase in FS% than the summation of the individual effects. Figure 4.4.c and Figure 4.4.d show the cross section curves of Figure 4.4.a. The result shows that for small cog.A<sub>0</sub>, the effect of  $k_r$  is not important (Figure 4.4.c), *i.e.* the effect of E-site tRNA release is less important if there is hungry codon in the A-site. Therefore, the model suggests that in the presence of a slippery P-site (high  $k_s$ ) with no hungry codon effect (large cog.A<sub>0</sub>), a higher rate of E-site tRNA release can still result in a higher FS%. In contrast, for lower rates of E-site tRNA release, the model predicts substantial FS% in the presence of P-site slippery sites and hungry codons (Figure 4.4.d).

#### 4.6.2 Empirical studies

To understand the importance of the release of E-site tRNA on +1 PRF, an *in vivo* dual fluorescence reporter system in *E. coli* is used to study the effect of the E-site stability on +1 PRF (see Materials and Methods). The reporter system (Figure 4.5.a) allows measurement of frameshift efficiency for different recoding sites by calculating the ratio of green to red fluorescence. All possible E-site codons (13 sense codon = 16 potential codons – 3 stop codons) in the RF2 frameshift site have been tested under the condition that SD-like sequence was kept intact. Statistical analysis was applied to all datasets according to Jacobs *et al.*, 2004 [28]. Ten replicates for the mutants and twenty replicates for the control were performed to satisfy the minimum sample requirement. The standard error for FS% for different mutants was less than 2 %. The results show that the presence of an A:U pair in the second position of the E-site codon in the RF2 frameshift site results in higher frameshifting as compared to a G:C pair in the same position (Figure 4.5.b). Importantly, frameshift efficiency can be

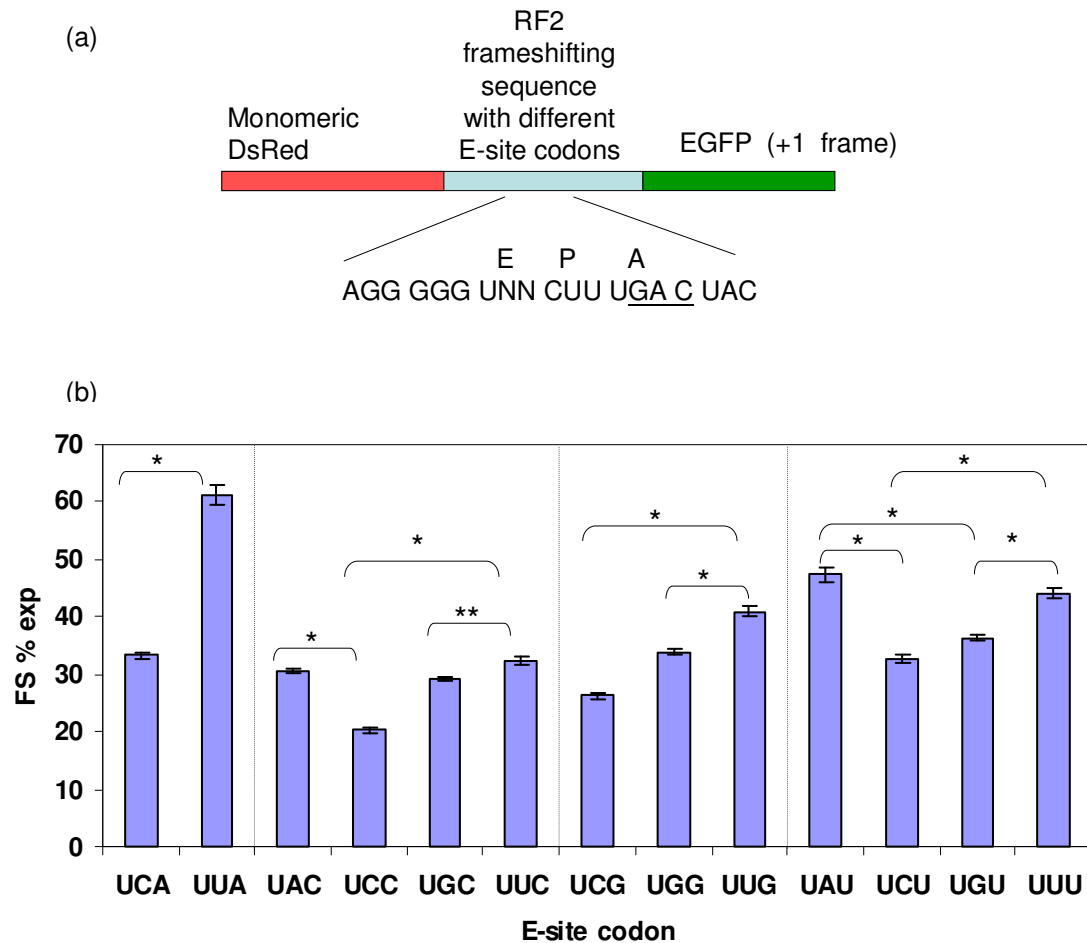


Figure 4.5. The effect of different E-site codon:anticodon interactions on frameshift efficiency. (a) The sequence design of the dual fluorescence reporter system. The E, P and A denote the codon in the E-, P-, and A-sites when +1 frameshifting is taking place. The +1 frame A-site codon is underlined. (b) Experimentally obtained frameshift efficiency for different E-site codons. The error bars indicate the standard deviation. \* indicates significant difference ( $p < 0.001$ ). \*\* indicates significant difference ( $p < 0.01$ ).

generally categorized into three levels based on the number of hydrogen bonds in the base pair interaction (Table 4.1). A:U pairs in both the second and the third positions have the lowest number of hydrogen bonds and promote the highest frameshifting. One A:U pair and one C:G pair in the second and the third positions result in the intermediate level of frameshifting. With the highest number of hydrogen bonds, C:G pairs in both the second and the third positions result in the lowest frameshifting. An interestingly unexpected result is UGG as an E-site codon. The frameshift efficiency for UGG in the E-site is comparable to that for one G:C and one A:U base pairs in the second and third positions in the E-site. This observation may result from factors not accounted for in the model or perhaps be a result of the reporter protein. The absence of modified nucleoside pseudouridine ( $\Psi$ ) at position 38-40 in tRNA<sup>Trp</sup><sub>CCA</sub> could also be a reason for less efficient binding of this tRNA to the E-site. It is suggested that deficiency of modified nucleosides may change tRNA structure, resulting in different ribosome:tRNA interactions [29]. However, the exact reason for relatively higher FS% for UGG in the E-site is not known. These FS% data are less likely to be due to the availability of tRNA for the specific codon in the E-site, because we observed no obvious correlation between FS% and tRNA concentration for the E-site codons (Figure 4.S1 in Supplementary Data). These experimental observations emphasize the effect of E-site stability on +1 PRF, which is consistent with the computational simulations described above.

#### 4.6.3 Parameter estimation

The rate constant for the P-site tRNA slippage ( $k_s$ ) can be estimated by combining the kinetic model and the experiments results. Changing E-site stabilities by using different E-site codons while maintaining the identity of the P-site codon enables manipulation of  $k_r$  at a constant  $k_s$ .  $k_r$  is assumed to be a function of stimulatory

Table 4.1 Three levels of +1 frameshift efficiency for different E-site codons

		3rd Base			
		A	U	C	G
2nd Base	A		47	31	
	U	61	44	32	41
	C	33	33	20	26
	G		36	29	34

- Highest frameshifting: A:U basepairs at both 2nd and 3rd positions of the E-site codon
- Intermediate frameshifting: One A:U and one G:C at the 2nd and 3rd position of the E-site codon
- Lowest frameshifting: G:C basepairs at both 2nd and 3rd positions of the E-site codon

signals, tRNA:mRNA (codon:anticodon) and tRNA:ribosome interactions in the E-site (see Materials and Methods). Chi-square analyses were performed to obtain optimum values for  $k_s$  and  $k_r$ , which give the best fit of the model predictions and the experimental results. Figure 4.6 shows the data-fitting result of the model prediction (solid line) and the experimentally detected FS% (diamonds).  $k_s$  was determined to be  $1.9 \text{ s}^{-1}$  for the RF2 frameshifting sequence and parameters for calculating  $k_r$  are listed in Table 4.S4 in Supplementary Data. Modifying factors were used to account for other factors (e.g. tRNA:ribosome interactions, base modification) that may affect the contribution of the base pairing on  $k_r$ . The modifying factor for tRNA<sub>QTA</sub><sup>Tyr</sup> was observed to be 2.18. The value is consistent with the observation that the binding efficiency of Q34-tRNA<sup>Tyr</sup> to triplet programmed ribosomes is two fold more than G34-tRNA<sup>Tyr</sup> [30]. Modifying factors for other tRNAs are less than one, which may suggest that for tRNA<sup>Phe</sup>, tRNA<sup>Leu</sup>, tRNA<sup>Ser</sup>, tRNA<sup>Cys</sup>, and tRNA<sup>Trp</sup>, other interactions in the E-site could reduce the contribution of codon:anticodon interactions on  $k_r$ . The correlation between FS%<sub>exp</sub> and free energy change of the codon:anticodon interactions in the E-site is shown in Figure 4.7.a and the correlation between FS%<sub>exp</sub> and apparent E-site stability (free energy change of codon:anticodon interactions in the E-site multiplied by modifying factors) is shown in Figure 4.7.b. Importantly, frameshift efficiency is observed to inversely correlate with the E-site stability and this observation is more clear when codon:anticodon interactions and other interactions in the E-site are all considered.

## 4.7 Discussion

### 4.7.1 Comparison of the three models

Figure 4.1 presents three possible pathways for ribosomes to synthesize +1 frameshift proteins in *E. coli*. We believe that all three pathways can occur *in vivo* but that



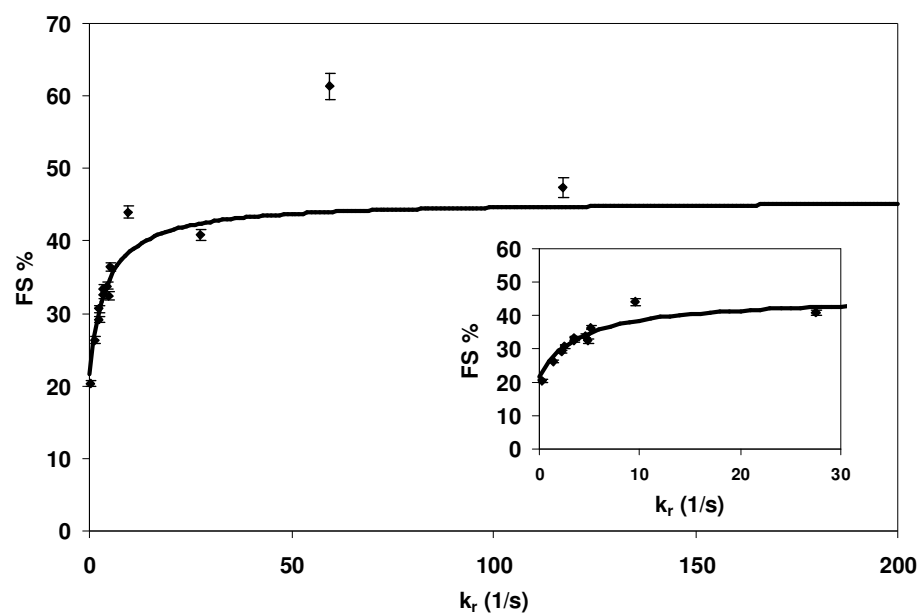


Figure 4.6. Data fit of frameshift efficiency for different codon:anticodon interactions in the E-site. By using chi-square analysis, the optimum value for  $k_s$  is  $1.9 \text{ s}^{-1}$  for RF2 frameshifting sequence. The diamonds show the experimentally obtained frameshift efficiency for different E-site codons ( $k_r = A' \exp(-m\Delta G_c/RT)$ , see Materials and Methods). The solid line indicates model predicted FS% for different  $k_r$  at  $k_s = 1.9 \text{ s}^{-1}$ .

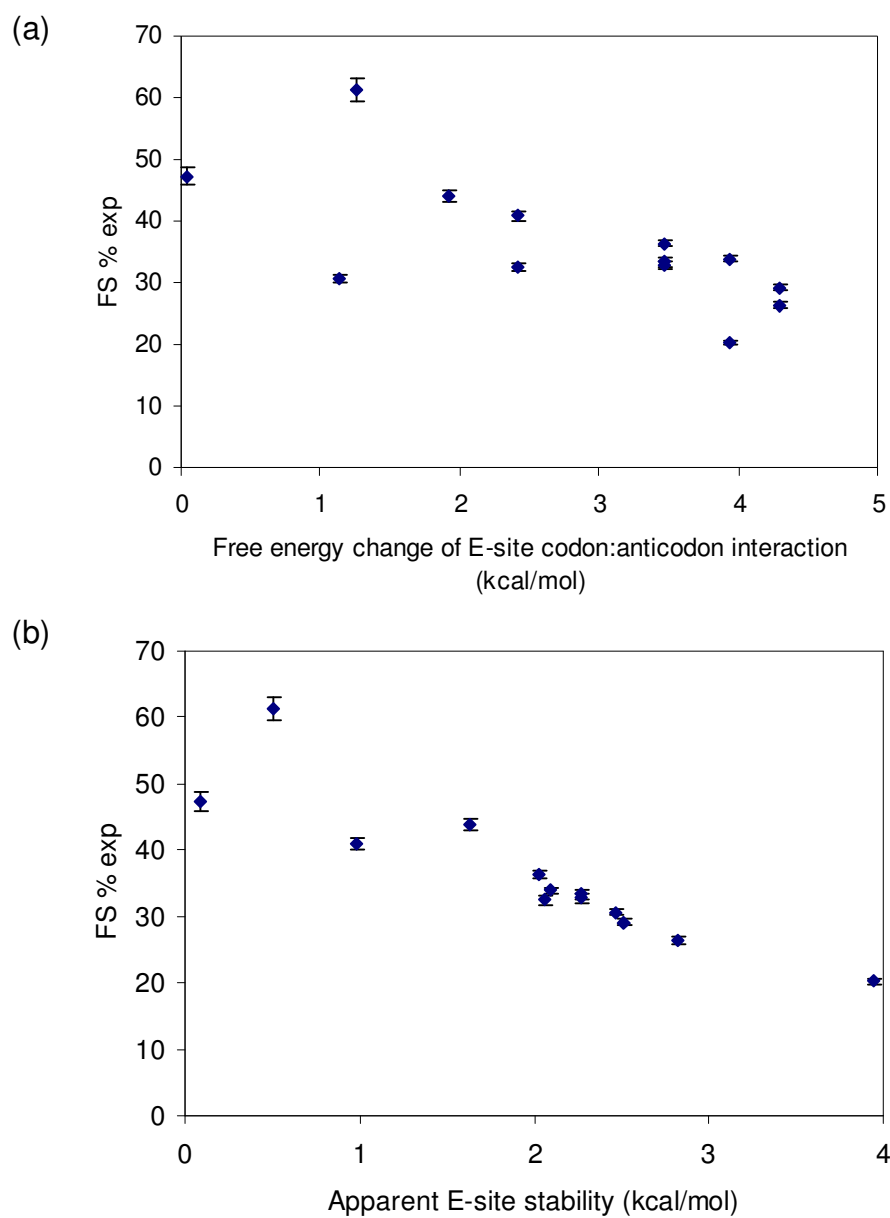


Figure 4.7. (a) The correlation between FS%<sub>exp</sub> and the free energy change of the E-site codon:anticodon interactions. (b) The correlation between FS%<sub>exp</sub> and the apparent E-site stability obtained by free energy change of the E-site codon:anticodon interactions multiplied by modifying factors (Table 4.S4).

Model 2 is the dominant pathway for +1 PRF. Model 1 involves simultaneous slippage of E- and P-site tRNAs to the +1 frame, an energetically unfavorable process less likely to occur. To test this hypothesis, the effect of  $k_{se}$  (the rate constant that determines if the ribosome complex would get into Model 1) was studied and it was found that FS% remained at a similar level at different values of  $k_{se}$  (Figure 4.S2 in Supplementary Data). A recently published study observed no correlation between +1 PRF efficiency and the stability of complex of E-site tRNA base pairing with +1 frame [18]. Similarly, the effect of  $k_{sd}$ , which governs whether the ribosome complex enters Model 1A pathway, was also studied. The effect of  $k_{sd}$  is observed to be less significant on FS% (Figure 4.S3 in Supplementary Data). These observations suggest that Model 1 and Model 1A may contribute much less than Model 2 or 3 to the overall FS%. In Model 3, the formation of  $P_0$ , the major precursor of frameshift products, depends on aa-tRNA rejection. The aa-tRNA that reaches the accommodation step is more likely to be cognate, because it has already passed through the selective codon recognition and GTPase activation steps. As a result, this aa-tRNA is less likely to be rejected, and thus the probability for ribosomes to form  $P_0$  is less. Although in Model 2, formation of  $P_0$  also depends on dissociation of aa-tRNA, the reversible nature of the codon recognition step and higher concentration of non- and near-cognate tRNAs relative to cognate tRNA, together with same rates for forward reactions of codon recognition for both substrates, make  $P_0$  formation likely. Therefore, Model 2 would result in formation of a significantly larger fraction of the  $P_0$  state ribosomes as compared to Model 3. Thus, we propose that Model 2 is the major pathway for +1 PRF *in vivo*.

#### 4.7.2 Role of the E-site

The function of ribosome E-site is still under debate in the literature. Some studies

suggest the E-site interactions are functionally important for maintaining the reading frame [16,31,32], while others suggest the E-site tRNA binds to the ribosome in a labile manner [33,34]. The results presented in this study are fundamentally helpful to explain different E-site effects suggested by different studies. In the proposed mechanism,  $k_r$  represents the de-occupation of E-site tRNA. Our model results show that the effect of E-site interactions on +1 PRF is more significant when  $k_r$  is smaller than  $10 \text{ s}^{-1}$  and the effect is less when  $k_r$  is in a range of larger values (Figure 4.2.c and Figure 4.4.c). We believe that different views of the E-site function can be due to the result of different experimental conditions, which produce different ranges of  $k_r$ . It has been suggested that the ionic conditions, physical parameters (pH, temperature, etc.), and material preparation methods all affect the binding affinity of tRNA to the ribosome [35]. Therefore, for buffer conditions or mRNA sequences for which  $k_r$  is in the range of smaller values, the effect of E-site interactions on +1 frameshift efficiency can be observed [16-18,31,32]. On the other hand, for buffer conditions or mRNA sequences for which  $k_r$  is in the range of larger values, the effect of E-site interactions is less important for translation elongation [33]. This observation clearly demonstrates the utility of a modeling approach to help reconcile disparate observations from the literature.

A question may remain: which range of  $k_r$  should be expected? The data fitting (Figure 4.6) shows that the range of  $k_r$  is actually different for different E-site codons *in vivo*. For E-site codons UAU, UUA, UUG and UUU,  $k_r$  are in a range of large values ( $\geq 10 \text{ s}^{-1}$ ). For all the other tested E-site codons in the present study, the  $k_r$  values are smaller ( $< 10 \text{ s}^{-1}$ ). Previous *in vitro* studies used polyU programmed ribosome to study E-site interactions [33,34], which could be the reason that smaller tRNA binding affinities to the ribosome E-site were observed. On the other hand,

weaker interactions in the ribosome E-site have been shown to reduce translational fidelity *in vivo* [17,18,32]. In the present study, the experiments support the importance of E-site interactions in +1 PRF in *E. coli* (Figure 4.5.b). The data show that an A:U base pairing in the E-site, which contains one less hydrogen bond than a G:C base pairing, results in higher frameshift efficiency. A recently published study using a monocistronic reporter system also showed that RF2 programmed frameshifting is inversely correlated with the E-site stability [18]. Taken together, the experimental data in the present study and the study by Sanders *et al.* [18] provide independent evidence that different E-site interactions may result in different ranges of  $k_r$  *in vivo*, illustrating the role of E-site stability on +1 PRF.

Mechanistically,  $k_r$  may be a function of mRNA:tRNA and tRNA:ribosome interactions at the E-site, stimulatory signals (SD sequence, mRNA structures, etc.), and spacing between the stimulatory signals and the E-site. The experimental results in the present study suggest that tRNA:mRNA base pairing in the E-site could be functionally important, supporting the X-ray crystal structures [22,23]. Previous experimental observations also support the effect of stimulatory factors and spacing on frameshifting. For example, it has been proposed that the interaction between the SD and anti-SD sequence in *E. coli prfB* mRNA precludes the binding of the E-site tRNA and therefore might facilitate destabilization of the E-site tRNA [16]. That study also showed that the spacing between the SD sequence and the frameshift site is critical for high frameshift efficiency. Mutations in the SD sequence have also been shown to cause significant reductions in frameshift efficiency [10]. In our model, the SD:antiSD interaction may play its role in RF2 frameshifting in *E. coli* in two ways. First, the presence of an SD:antiSD interaction enhances the release of E-site tRNA. As for the data fitting in this study, the rate constant for E-site tRNA release is assumed as

$k_r = A' \exp(m\Delta G_c/RT)$ . The presence of an SD-like sequence will result in a larger  $A'$  and therefore result in a higher rate of E-site tRNA release, paving the way for +1 PRF in *E. coli* as described in Model 2. Secondly, the SD:antiSD interaction may destabilize the ribosome complex, yielding unstable complex  $E'_0P'_0$ , which can directly interact with +1 frame aa-tRNA as described in Model 1A.

Stimulatory elements have also been found in the Ty3 and OAZ1 +1 PRF signals in yeast, and their effects also depended on strict spacing from the sites of frameshifting [6,36]. However, there is not yet any direct experimental evidence demonstrating the effect of E-site destabilization in Ty1 and Ty3 frameshifting. The prokaryotic ribosomal structure suggests that although there is no direct contact between E-site tRNA and P-site tRNA in the ribosome, the E-site tRNA might interact indirectly with the P-site tRNA through the 16S rRNA [37-39]. In agreement with these observations, our model of +1 PRF suggests that E-site tRNA dissociation might destabilize the mRNA ribosome interactions and affect the P-site tRNA slippage. Thus, ribosomes with an empty E-site may be more prone to slip.

#### 4.7.3 Role of the P-site

The computational modeling shows that for small values of  $k_s$  ( $k_s=0.05 \text{ s}^{-1}$ ), the effects of hungry codons in the A-site, and of rates of E-site tRNA release on FS% are less significant, thus demonstrating that P-site tRNA slippage is the dominant factor for +1 PRF in *E. coli*. As illustrated in Figure 4.1, the efficiency of +1 PRF is determined by two competing reaction branches: 1) zero-frame aa-tRNA selection followed by peptidyl transfer (PT), and 2) P-site tRNA slippage to the +1 frame, which is subsequently trapped by aa-tRNA selection and PT. The rate constant of slippage,  $k_s$ , depends on the stability of the  $P_0$  and  $P_1$  states. In other words, the analysis presented

here indicates that the less stable  $P_0$  and more stable  $P_1$  (which gives higher  $k_s$ ) should result in higher FS%. This is consistent with the previous experimental observations. Curran [9] showed that among 32 polynucleotides differing only in their P-site, tRNAs that form more cognate interactions with the +1 frame in the P-site had a 1000-fold increase in frameshift proteins than tRNAs mispairing with the +1 frame. Other factors such as wobble base modification and tRNA hypomodification have been shown to weaken base-pairing and stimulate tRNA slippage in the P-site [29,40-41]. It has also been suggested that features of tRNA structure outside of the anticodon contribute to the P-site stability and the ability to shift reading frames [42-44]. Moreover, in yeast, a mutant form of ribosomal protein L5 (*RPL5*) that promoted decreased ribosomal affinity for peptidyl-tRNA also promoted increased +1 PRF at a *TyI* signal [45].

The rate constant for P-site tRNA slippage has not been previously reported in the literature. Our kinetic model combined with experiments using different E-site interactions provides an approach to estimate  $k_s$ . Fitting the experimental data for RF2 frameshifting sequence (CUU U sequence in the P-site) yielded a rate constant of slippage  $\approx 1.9 \text{ s}^{-1}$ . The small magnitude of  $k_s$ , as compared to other rate constants in the model, is consistent with the idea that the slippage is the rate-limiting reaction in the +1 PRF mechanism.

#### **4.7.4 Role of the A-site**

Our model suggests that in the presence of a slippery P-site, a low availability of cognate aa-tRNA for zero frame (cog.A<sub>0</sub>) can enhance FS% by about two fold (Figure 4.3.a and 3C). A low concentration of cognate tRNA at the A-site (hungry codon) has been experimentally observed to promote frameshifting [46], consistent with the

model. We believe that the low availability of the zero frame cognate aa-tRNA (cog. $A_0$ ) can affect +1 PRF in two ways. First, the low availability of cog. $A_0$  slows down translation, which allows more time for the kinetically-driven +1 PRF event to take place. Secondly, the low availability of cog. $A_0$  increases the chance for the near-cognate tRNA to bind to the A-site. During the elongation cycle, both cognate and near-cognate tRNAs compete for the A-site. Since a near-cognate tRNA has more chances to be rejected after the codon recognition step or GTP hydrolysis step than the cognate tRNA, a low concentration of cognate tRNA is more likely to result in the ribosomes containing only P-site tRNA ( $P_0$ ), thus enhancing the probability of slippage. In support of this, studies in yeast show that mutants that only affect A-site affinities for aa-tRNAs do not affect +1 PRF efficiency [47,48].

#### **4.7.5 +1 PRF in eukaryotes**

The rate constants used in this study are based on data obtained using *E. coli* ribosomes. The finding of synergistic effects among E-, P-, and A-site interactions on +1 PRF is likely to be applicable to Ty1 expression in yeast and antizyme expression in mammalian cells. However, owing to differences in aa-tRNA abundance and ribosome structures between prokaryotes and eukaryotes, eukaryotic +1 PRF signals were not tested in the present study. For Ty3 frameshifting in yeast, it is suggested that a special P-site interaction may interfere with the binding of in-frame aa-tRNA and stabilize out-of-frame decoding [6]. According to our model, this observation suggests the possibility that a special tRNA interaction in the P-site may change  $k_{1s}$ . It is also likely that the Ty3 mechanism includes another reaction pathway for  $P_0$  to directly interact with a +1 frame aa-tRNA ternary complex. We believe that a quantitative kinetic model, similar to our current model, can be built for Ty3 frameshifting in yeast to understand this unique frameshifting process better.



#### 4.8 Supplementary data

From the mechanism (Figure 4.1), the formation rate of each component can be written as the following:

$$\frac{d[P_1A_1]}{dt} = k_{7s}[P_1A_1^*]$$

$$\frac{d[P_1A_1^*]}{dt} = k_{5s}[P_1A_1^{GDP}] - (k_{7s} + k_{6s})[P_1A_1^*]$$

$$\frac{d[P_1A_1^{GDP}]}{dt} = k_{4s}[P_1A_1^{GTPase^*}] - k_{5s}[P_1A_1^{GDP}]$$

$$\frac{d[P_1A_1^{GTPase^*}]}{dt} = k_{3s}[P_1A_1^{CR}] - k_{4s}[P_1A_1^{GTPase^*}]$$

$$\frac{d[P_1A_1^{CR}]}{dt} = k_{2s}[P_1A_1^i] - (k_{3s} + k_{-2s})[P_1A_1^{CR}]$$

$$\frac{d[P_1A_1^i]}{dt} = k_{1s}[A_1]([P_1] + [E_1P_1] + [E'_0P'_0]) + k_{-2s}[P_1A_1^{CR}] - (3k_{-1s} + k_{2s})[P_1A_1^i]$$

$$\frac{d[E_1P_1]}{dt} = k_{se}[E_0P_0] + k_{-1s}[P_1A_1^i] - (k_{-se} + k_{1s}[A_1])[E_1P_1]$$

$$\frac{d[E'_0P'_0]}{dt} = k_{sd}[E_0P_0] + k_{-1s}[P_1A_1^i] - (k_{-sd} + k_{1s}[A_1])[E'_0P'_0]$$

$$\frac{dP_1}{dt} = k_s[P_0] + k_{-1s}[P_1A_1^i] + k_{6s}[P_1A_1^*] - (k_{-s} + k_{1s}[A_1])[P_1]$$

$$\frac{d[P_0A_0]}{dt} = k_7[P_0A_0^*]$$

$$\frac{d[P_0A_0^*]}{dt} = k_5[P_0A_0^{GDP}] - (k_7 + k_6)[P_0A_0^*]$$

$$\frac{d[P_0A_0^{GDP}]}{dt} = k_4([E_0P_0A_0^{GTPase^*}] + [P_0A_0^{GTPase^*}]) - k_5[P_0A_0^{GDP}]$$

$$\frac{d[E_0P_0A_0^{GTPase}]}{dt} = k_3[E_0P_0A_0^{CR}] - k_4[E_0P_0A_0^{GTPase*}]$$

$$\frac{d[P_0A_0^{GTPase}]}{dt} = k_3[P_0A_0^{CR}] - k_4[P_0A_0^{GTPase*}]$$

$$\frac{d[E_0P_0A_0^{CR}]}{dt} = k_2[E_0P_0A_0^i] - (k_3 + k_{-2} + k_{r1})[E_0P_0A_0^{CR}]$$

$$\frac{d[P_0A_0^{CR}]}{dt} = k_{r1}[E_0P_0A_0^{CR}] + k_2[P_0A_0^i] - (k_3 + k_{-2})[P_0A_0^{CR}]$$

$$\frac{d[E_0P_0A_0^i]}{dt} = k_1[E_0P_0][A_0] + k_{-2}[E_0P_0A_0^{CR}] - (k_{-1} + k_2)[E_0P_0A_0^i]$$

$$\frac{d[P_0A_0^i]}{dt} = k_1[A_0][P_0] + k_{-2}[P_0A_0^{CR}] - (k_{-1} + k_2)[P_0A_0^i]$$

$$\frac{dP_0}{dt} = k_6[P_0A_0^*] + k_{-1}[P_0A_0^i] + k_{-s}[P_1] - (k_1[A_0] + k_s)[P_0]$$

By assuming steady state, the formation rates of the intermediates equal to zero. The expressions of non-frameshifted ( $P_0A_0$ ) and frameshift proteins ( $P_1A_1$ ) in terms of  $E_0P_0$  are solved by Matlab V7.2 (MathWorks Inc., USA). Frameshift efficiency has been defined as:

$$\text{Frameshift Efficiency (FS \%)} = \frac{P_1A_1}{P_0A_0 + P_1A_1} \times 100\%$$

Table 4.S1 The rate constants for different steps at 20°C. The rate constants used in the model equal to the rate constant at 20 °C times the fold change from 20°C to 37°C (Table 4.S2).

Step	Symbols	Rate constants(s <sup>-1</sup> )	
		Cognate	Near-cognate
Initial binding	k <sub>1</sub> , k <sub>1s</sub>	110 <sup>a,b</sup>	110 <sup>a,b</sup>
	k <sub>-1</sub> , k <sub>-1s</sub>	25 <sup>a</sup>	25 <sup>a</sup>
Codon Recognition	k <sub>2</sub> , k <sub>2s</sub>	100 <sup>a</sup>	100 <sup>a</sup>
	k <sub>-2</sub> , k <sub>-2s</sub>	0.2 <sup>a</sup>	80 <sup>a</sup>
GTPase activation and GTP hydrolysis*	k <sub>3</sub> , k <sub>4</sub> k <sub>3s</sub> , k <sub>4s</sub>	260 <sup>a</sup>	0.4 <sup>a</sup>
EF-Tu conformational change (dissociation)	k <sub>5</sub> , k <sub>5s</sub>	60 <sup>a</sup>	70 <sup>a</sup>
tRNA rejection	k <sub>6</sub> , k <sub>6s</sub>	0.3 <sup>a</sup>	6 <sup>a</sup>
Accommodation	k <sub>7</sub> , k <sub>7s</sub>	7 <sup>a</sup>	0.1 <sup>a</sup>
E-tRNA, P-tRNA slippage	k <sub>se</sub> ,	0.01 <sup>c</sup>	0.01 <sup>c</sup>
	k <sub>-se</sub>	0.01 <sup>c</sup>	0.01 <sup>c</sup>
E-tRNA, P-tRNA destabilization	k <sub>sd</sub>	1 <sup>c</sup>	1 <sup>c</sup>
	k <sub>-sd</sub>	100 <sup>c</sup>	100 <sup>c</sup>
P-tRNA slippage	k <sub>s</sub>	0.05-5 <sup>d</sup>	0.05-5 <sup>d</sup>
	k <sub>-s</sub>	5 <sup>c</sup>	5 <sup>c</sup>
E-tRNA release	k <sub>r</sub>	1-100 <sup>d</sup>	1-100 <sup>d</sup>

<sup>a</sup> Rodnina *et al.*, 2005 [19].

<sup>b</sup> μM<sup>-1</sup>s<sup>-1</sup>

<sup>c</sup> Assumed values in the model

<sup>d</sup> Assumed range of values in the model

Table 4.S2 The activation energy for different steps in the model and the fold change of the rate constants ( $k_{310K}/k_{293K}$ , from 20°C to 37°C).

	$E_a$ (kJ/mol)	$k_{310K}/k_{293K}$
$k_1, k_{1s}$	$10 \pm 6^a$	1.25
$k_{-1}, k_{-1s}$	$46 \pm 5^a$	2.82
$k_2, k_{2s}$	$38 \pm 8^a$	2.36
$k_{2c}, k_{2s}$	$44 \pm 5^a$	2.7
$k_{2nc}$	$38^e$	2.36
$k_{3c}, k_{3s}$	$55^b$	3.45
$k_{3nc}$	$55^b + 15^c$	4.85
$k_{4c}, k_{4s}$	$55^b$	3.45
$k_{4nc}$	$55^b + 15^c$	4.85
$k_{5c}, k_{5s}$	$155^d$	33
$k_{5nc}$	$155^d$	33
$k_{6c}, k_{6s}$	$55^e$	3.45
$k_{6nc}$	$55^e - 8^e$	2.88
$k_{7c}, k_{7s}$	$55^e$	3.45
$k_{7nc}$	$55^e + 8^c$	4.13

<sup>a</sup> Rodnina et al. 1996 [49].

<sup>b</sup> Thompson et al. 1980 [50].

<sup>c</sup> Gromadski et al. 2006 [51].

<sup>d</sup> Karim et al. 1986. [52]

<sup>e</sup> Assumed values in the model

Table 4.S3 The concentration of the components used in the model

Components	Symbols	Parameter values
Initial reactant	$E_0P_0$	1 <sup>a</sup>
Cognate zero-frame aa-tRNA	cog. $A_0$ (%)	0.1-10 <sup>b</sup>
Near cognate zero frame aa-tRNA	nc. $A_0$ (%)	19.9-10 <sup>b</sup>
+1 frame aa-tRNA <sup>b</sup>	$A_1$ (%)	10 <sup>c</sup>

<sup>a</sup> Assumed value in the model ( $\mu$ M).

<sup>b</sup> Assumed range of values in the model. We assume that the total concentration of the near cognate aa-tRNA and the cognate aa-tRNA for a particular codon equals to 20% of the total aa-tRNA pool, i.e. nc. $A_0$  + cog. $A_0$  = 20%.

<sup>c</sup> Assumed value in the model. We assume that the +1 frame A-site is always a non hungry codon and the corresponding aa-tRNA is always cognate.

Table 4.S4 Optimum parameter values for calculating  $k_r$  obtained from data fitting.  $k_r$  is calculated as  $k_r = A' \exp(-m\Delta G_c/RT)$ . From data fitting, pre-exponential factor  $A'$  is obtained to be  $135 \text{ s}^{-1}$  and modifying factors are listed as the following.

E-site codon	tRNA	Modifying factor	Parameter values
UAU	tRNA <sub>QTA</sub> <sup>Tyr</sup>	m1	2.18
UAC	tRNA <sub>QTA</sub> <sup>Tyr</sup>	m1	2.18
UCA	tRNA <sub>cmo</sub> <sup>5 UGA</sup> <sup>Ser</sup>	m2	0.66
UCC	tRNA <sub>GGA</sub> <sup>Ser</sup>	-	1.00 <sup>a</sup>
UCG	tRNA <sub>CGA</sub> <sup>5 Ser</sup> tRNA <sub>cmo</sub> <sup>5 UGA</sup> <sup>Ser</sup>	m2	0.66
UCU	tRNA <sub>GGA</sub> <sup>5 Ser</sup> tRNA <sub>cmo</sub> <sup>5 UGA</sup> <sup>Ser</sup>	m2	0.66
UUA	tRNA <sub>UAA</sub> <sup>Leu</sup>	m3	0.41
UUG	tRNA <sub>CAA</sub> <sup>Leu</sup>	m3	0.41
UUU	tRNA <sub>GAA</sub> <sup>Phe</sup>	m4	0.85
UUC	tRNA <sub>GAA</sub> <sup>Phe</sup>	m4	0.85
UGU	tRNA <sub>GCA</sub> <sup>Cys</sup>	m5	0.59
UGC	tRNA <sub>GCA</sub> <sup>Cys</sup>	m5	0.59
UGG	tRNA <sub>CCA</sub> <sup>Trp</sup>	m6	0.53

<sup>a</sup>UCC is read by tRNA without modification at either position 34 or position 37. Thus we assume modifying factor for tRNA<sub>GGA</sub><sup>Cys</sup> equals 1.

#### **4.9 Conclusion**

A detailed kinetic model for +1 PRF in *E. coli* has been presented and the effect of E-site stabilities on +1 PRF has been experimentally demonstrated. According to the model results, a combination of stimulatory signals leading to the release of deacylated tRNA in the E-site, tRNA slippage in the P-site, and the hungry codon effect in the A-site synergistically promote efficient +1 ribosomal frameshifting. The experimental result suggested that weaker codon:anticodon interactions in the E-site correlate with higher +1 PRF efficiency in *E. coli*. Our mathematical analysis shows that the rate of P-site tRNA slippage is the dominant factor, while the effect of hungry codon in the A-site and E-site tRNA destabilization further enhance +1 PRF. We propose that E-site empty ribosomes, which facilitate the P-site tRNA slippage, is the driving force for +1 PRF.

#### **4.10 Acknowledgments**

We are thankful to Navneetha Santhanam, Dr. Fernando Escobedo, and Dr. Abraham Stroock for their insightful comments and critiques of this work. We gratefully acknowledge Dr. Matthew DeLisa for the DsRed gene sequence. We also acknowledge Robert Kuczenski for advice in developing the Matlab program.

## REFERENCES

1. Craigen, W.J. and Caskey, C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, **322**, 273-275.
2. Belcourt, M.F. and Farabaugh, P.J. (1990) Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell*, **62**, 339-352.
3. Farabaugh, P.J., Zhao, H. and Vimaladithan, A. (1993) A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell*, **74**, 93-103.
4. Asakura, T., Sasaki, T., Nagano, F., Satoh, A., Obaishi, H., Nishioka, H., Imamura, H., Hotta, K., Tanaka, K., Nakanishi, H., *et al.* (1998) Isolation and characterization of a novel actin filament-binding protein from *Saccharomyces cerevisiae*. *Oncogene*, **16**, 121-130.
5. Morris, D.K. and Lundblad, V. (1997) Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr. Biol.*, **7**, 969-976.
6. Palanimurugan, R., Scheel, H., Hofmann, K. and Dohmen, R.J. (2004) Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J.*, **23**, 4857-4867.
7. Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F. and Hayashi, S. (1995) Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell*, **80**, 51-60.
8. Lindsley, D. and Gallant, J. (1993) On the directional specificity of ribosome frameshifting at a "hungry" codon. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 5469-5473.
9. Curran, J.F. (1993) Analysis of effects of tRNA:Message stability on frameshift frequency at the *Escherichia coli* RF2 programmed frameshift site. *Nucleic Acids Res.*, **21**, 1837-1843.
10. Weiss, R.B., Dunn, D.M., Dahlberg, A.E., Atkins, J.F. and Gesteland, R.F. (1988) Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.*, **7**, 1503-1507.
11. Harger, J.W., Meskauskas, A. and Dinman, J.D. (2002) An "integrated model" of programmed ribosomal frameshifting. *Trends Biochem. Sci.*, **27**, 448-454.
12. Farabaugh, P.J. and Bjork, G.R. (1999) How translational accuracy influences reading frame maintenance. *EMBO J.*, **18**, 1427-1434.



13. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2004) P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221-230.
14. Sundararajan,A., Michaud,W.A., Qian,Q., Stahl,G. and Farabaugh,P.J. (1999) Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. *Mol. Cell*, **4**, 1005-1015.
15. Kawakami,K., Pande,S., Faiola,B., Moore,D.P., Boeke,J.D., Farabaugh,P.J., Strathern,J.N., Nakamura,Y. and Garfinkel,D.J. (1993) A rare tRNA-arg(CCU) that regulates TyI element ribosomal frameshifting is essential for TyI retrotransposition in *Saccharomyces cerevisiae*. *Genetics*, **135**, 309-320.
16. Marquez,V., Wilson,D.N., Tate,W.P., Triana-Alonso,F. and Nierhaus,K.H. (2004) Maintaining the ribosomal reading frame: The influence of the E site during translational regulation of release factor 2. *Cell*, **118**, 45-55.
17. Sergiev,P.V., Lesnyak,D.V., Kiparisov,S.V., Burakovsky,D.E., Leonov,A.A., Bogdanov,A.A., Brimacombe,R. and Dontsova,O.A. (2005) Function of the ribosomal E-site: A mutagenesis study. *Nucleic Acids Res.*, **33**, 6048-6056.
18. Sanders,C.L. and Curran,J.F. (2007) Genetic analysis of the E site during RF2 programmed frameshifting. *RNA*, **13**, 1483-1491.
19. Rodnina,M.V., Gromadski,K.B., Kothe,U. and Wieden,H.J. (2005) Recognition and selection of tRNA in translation. *FEBS Lett.*, **579**, 938-942.
20. Nierhaus,K.H. (1990) The allosteric three-site model for the ribosomal elongation cycle: Features and future. *Biochemistry*, **29**, 4997-5008.
21. Dinos,G., Kalpaxis,D.L., Wilson,D.N. and Nierhaus,K.H. (2005) Deacylated tRNA is released from the E site upon A site occupation but before GTP is hydrolyzed by EF-tu. *Nucleic Acids Res.*, **33**, 5291-5296.
22. Yusupova,G., Jenner,L., Rees,B., Moras,D. and Yusupov,M. (2006) Structural basis for messenger RNA movement on the ribosome. *Nature*, **444**, 391-394.
23. Jenner,L., Rees,B., Yusupov,M. and Yusupova,G. (2007) Messenger RNA conformations in the ribosomal E site revealed by X-ray crystallography. *EMBO Rep.*, **8**, 846-850.
24. Korostelev,A., Trakhanov,S., Laurberg,M. and Noller,H.F. (2006) Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, **126**, 1065-1077.
25. Selmer,M., Dunham,C.M., Murphy,F.V.,4th, Weixlbaumer,A., Petry,S., Kelley,A.C., Weir,J.R. and Ramakrishnan,V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935-1942.

26. Campbell,R.E., Tour,O., Palmer,A.E., Steinbach,P.A., Baird,G.S., Zacharias,D.A. and Tsien,R.Y. (2002) A monomeric red fluorescent protein. *PNAS*, **99**, 7877-7882.
27. Klump,H.H. (2006) Exploring the energy landscape of the genetic code. *Arch. Biochem. Biophys.*, **453**, 87-92.
28. Jacobs,J.L. and Dinman,J.D. (2004) Systematic analysis of bicistronic reporter assay data. *Nucleic Acids Res.*, **32**, e160.
29. Urbonavicius,J., Qian,Q., Durand,J.M., Hagervall,T.G. and Bjork,G.R. (2001) Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J.*, **20**, 4863-4873.
30. Noguchi,S., Nishimura,Y., Hirota,Y. and Nishimura,S. (1982) Isolation and characterization of an *Escherichia coli* mutant lacking tRNA-guanine transglycosylase. function and biosynthesis of queuosine in tRNA. *J. Biol. Chem.*, **257**, 6544-6550.
31. Nierhaus,K.H. (2006) Decoding errors and the involvement of the E-site. *Biochimie*, **88**, 1013-1019.
32. O'Connor,M., Willis,N.M., Bossi,L., Gesteland,R.F. and Atkins,J.F. (1993) Functional tRNAs with altered 3' ends. *EMBO J.*, **12**, 2559-2566.
33. Semenov,Y.P., Rodnina,M.V. and Wintermeyer,W. (1996) The "allosteric three-site model" of elongation cannot be confirmed in a well-defined ribosome system from *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 12183-12188.
34. Lill,R. and Wintermeyer,W. (1987) Destabilization of codon-anticodon interaction in the ribosomal exit site. *J. Mol. Biol.*, **196**, 137-148.
35. Schilling-Bartetzko,S., Franceschi,F., Sternbach,H. and Nierhaus,K.H. (1992) Apparent association constants of tRNAs for the ribosomal A, P, and E sites. *J. Biol. Chem.*, **267**, 4693-4702.
36. Li,Z., Stahl,G. and Farabaugh,P.J. (2001) Programmed +1 frameshifting stimulated by complementarity between a downstream mRNA sequence and an error-correcting region of rRNA. *RNA*, **7**, 275-284.
37. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883-896.
38. Ramakrishnan,V. and Moore,P.B. (2001) Atomic structures at last: The ribosome in 2000. *Curr. Opin. Struct. Biol.*, **11**, 144-154.

39. Spahn,C.M., Beckmann,R., Eswar,N., Penczek,P.A., Sali,A., Blobel,G. and Frank,J. (2001) Structure of the 80S ribosome from *Saccharomyces cerevisiae*-- tRNA-ribosome and subunit-subunit interactions. *Cell*, **107**, 373-386.
40. Yokoyama,S., Watanabe,T., Murao,K., Ishikura,H., Yamaizumi,Z., Nishimura,S. and Miyazawa,T. (1985) Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. *Proc. Natl. Acad. Sci. U. S. A.*, **82**, 4905-4909.
41. Tsuchihashi,Z. and Brown,P.O. (1992) Sequence requirements for efficient translational frameshifting in the *Escherichia coli dnaX* gene and the role of an unstable interaction between tRNA(lys) and an AAG lysine codon. *Genes Dev.*, **6**, 511-519.
42. Hansen,T.M., Baranov,P.V., Ivanov,I.P., Gesteland,R.F. and Atkins,J.F. (2003) Maintenance of the correct open reading frame by the ribosome. *EMBO Rep.*, **4**, 499-504.
43. Schultz,D.W. and Yarus,M. (1994) tRNA structure and ribosomal function. II. interaction between anticodon helix and other tRNA mutations. *J. Mol. Biol.*, **235**, 1395-1405.
44. Smith,D. and Yarus,M. (1989) Transfer RNA structure and coding specificity. I. evidence that a D-arm mutation reduces tRNA dissociation from the ribosome. *J. Mol. Biol.*, **206**, 489-501.
45. Meskauskas,A. and Dinman,J.D. (2001) Ribosomal protein L5 helps anchor peptidyl-tRNA to the P-site in *Saccharomyces cerevisiae*. *RNA*, **7**, 1084-1096.
46. Gallant,J. and Lindsley,D. (1993) Ribosome frameshifting at hungry codons: Sequence rules, directional specificity and possible relationship to mobile element behaviour. *Biochem. Soc. Trans.*, **21**, 817-821.
47. Meskauskas,A., Baxter,J.L., Carr,E.A., Yasenchak,J., Gallagher,J.E., Baserga,S.J. and Dinman,J.D. (2003) Delayed rRNA processing results in significant ribosome biogenesis and functional defects. *Mol. Cell. Biol.*, **23**, 1602-1613.
48. Meskauskas,A., Petrov,A.N. and Dinman,J.D. (2005) Identification of functionally important amino acids of ribosomal protein L3 by saturation mutagenesis. *Mol. Cell Biol.*, **25**, 10863-10874.
49. Rodnina,M.V., Pape,T., Fricke,R., Kuhn,L., Wintermeyer,W. (1996) Initial binding of the elongation factor tu.GTP.aminoacyl-tRNA complex preceding codon recognition on the ribosome. *J. Biol. Chem.*, **271**, 646-652.
50. Thompson,R.C., Dix,D.B., Eccleston,J.F. (1980) Single turnover kinetic studies of guanosine triphosphate hydrolysis and peptide formation in the elongation factor

tu-dependent binding of aminoacyl-tRNA to *Escherichia coli* ribosomes. *J. Biol. Chem.*, **255**: 11088-11090.

51. Gromadski, K.B., Daviter, T. and Rodnina, M.V. (2006) A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Mol. Cell* **21**, 369-377.
52. Karim, A.M. and Thompson, R.C. (1986) Guanosine 5'-O-(3-thiotriphosphate) as an analog of GTP in protein biosynthesis. The effects of temperature and polycations on the accuracy of initial recognition of aminoacyl-tRNA ternary complexes by ribosomes. *J. Biol. Chem.*, **261**, 3238-3243.

## CHAPTER 5

### FSSCAN: A MECHANISM-BASED PROGRAM TO IDENTIFY +1 RIBOSOMAL FRAMESHIFT HOT SPOTS

#### **5.1 Preface**

This chapter is adapted from Liao, P.Y., Choi, Y.S., Lee, K.H. 2009 FSscan: a mechanism-based program to identify +1 ribosomal frameshift hotspots. *Nucleic Acids Research*, doi:10.1093/nar/gkp796. Motivated by the result from Chapter 3, a bioinformatic program is developed to detect +1 programmed ribosomal frameshifting hotspots in the *Escherichia coli* genome. Candidate sequences identified by the program showed higher frameshift efficiency compared to a randomly design sequence *in vivo*.

#### **5.2 Abstract**

In +1 programmed ribosomal frameshifting (PRF), ribosomes skip one nucleotide toward the 3' end during translation. Most of the genes known to demonstrate +1 PRF have been discovered by chance or by searching homologous genes. Here, a bioinformatic framework called FSscan is developed to perform a systematic search for potential +1 frameshift sites in the *Escherichia coli* genome. Based on a current state of the art understanding of the mechanism of +1 PRF, FSscan calculates scores for a 16-nucleotide window along a gene sequence according to different effects of the stimulatory signals, and ribosome E-, P-, and A-site interactions. FSscan successfully identified the +1 PRF site in *prfB* and predicted *yehP*, *pepP*, *nuoE* and *cheA* as +1 frameshift candidates in the *E. coli* genome. Empirical results demonstrated that potential +1 frameshift sequences identified promoted significant levels of +1 frameshifting *in vivo*. Mass spectrometry analysis confirmed the presence of the

frameshift proteins expressed from a *yehP-egfp* fusion construct. FSscan allows a genome-wide and systematic search for +1 frameshift sites in *E. coli*. The results have implications for bioinformatic identification of novel frameshift proteins, ribosomal frameshifting, coding sequence detection, and the application of mass spectrometry on studying frameshift proteins.

### 5.3 Introduction

Translation is a highly accurate process. The frequency of decoding error is estimated to be on the order of  $10^{-5}$  per codon [1]. Programmed ribosomal frameshifting (PRF) is a coded shift in the reading frame during translation. Consequently, mRNAs with PRF features may yield two different protein products, an inframe product and a frameshift product. In +1 PRF, the ribosome skips over one nucleotide toward the 3' direction. As of September 2009, 88 cases of +1 PRF have been found in different organisms in the RECODE database [2]. +1 PRF has been observed to occur during the translation of *prfB* to produce release factor 2 (RF2) in *Escherichia coli* [3]. In *Saccharomyces cerevisiae* four retrotransposable elements, Ty1, Ty2, Ty3, and Ty4 [4-6], and three genes, *ABP140* [7], *EST3* [8], and *OAZ1* [9] use +1 PRF. The expression of mammalian antizyme has also been shown to involve +1 PRF [10].

A genome-wide prediction of +1 frameshift sites is currently a difficult task because the sequence elements for +1 frameshifting are diverse among the organisms. To date, most of the known genes involving +1 PRF have been discovered by chance, and in some cases, by searching homologous genes. Several computer programs have been developed to identify +1 frameshift sites [11-12]. Shah *et al.* [11] hypothesized that selective pressure would have rendered potential frameshift sites under-abundant in protein coding sequences. In this study, a computer program was developed to identify

oligos that are over- or under-represented for reasons other than codon bias. Their result suggested that the heptanucleotides CUU AGG C and CUU AGU U, +1 PRF sites for the production of *ABP140* and *EST3*, respectively, rank among the least represented of the heptanucleotides in the coding sequence of *S. cerevisiae*. While the approach is able to identify novel sequences, this method did not account for stimulatory signals. The program “FSFinder” by Moon *et al.* [12] used known components of a frameshift cassette for predicting both -1 and +1 PRF sites. This method achieves a high sensitivity and a high specificity (0.88 and 0.97, respectively) for predicting +1 PRF. However, FSFinder does not predict novel +1 frameshift sites in *E. coli*. A novel antizyme gene, whose expression requires +1 frameshifting, was found in the zebrafish *Danio rerio* by a protein BLAST search against the translated nucleotide database of the known antizyme family sequence [13]. While the method successfully identified novel genes requiring +1 frameshifting, the approach is limited to the antizyme family in eukaryotic cells.

Recently, a mathematical model revealed that destabilization of the deacylated tRNA in the ribosomal E-site, rearrangement of the peptidyl-tRNA in the ribosomal P-site, and availability of the cognate aminoacylated tRNA (aa-tRNA) corresponding to the ribosomal A-site act synergistically to promote efficient +1 PRF in *E. coli* [14]. Motivated by this result, one might identify potential +1 frameshift sites in the *E. coli* genome by searching sequences with a combination of stimulatory, E-, P-, and A-site features. In this study, FSscan is developed to perform a systematic and genome-wide search for potential +1 frameshift sites in *E. coli*. Based on a current state of art understanding of the mechanism of +1 PRF, FSscan looks for a 16-nucleotide sequence with possible synergistic effects in the *E. coli* genome. Potential +1 frameshift sequences so identified are shown to promote significant levels of +1

frameshifting *in vivo*. The mass spectrometry data obtained from a multiple reaction monitoring assay (MRM), a specific and sensitive mass scan method [15], experimentally confirms the expression of the predicted frameshift protein. Importantly, current methods of coding sequence detection generally do not take into account the shift of the reading frames and only a few algorithms assign a frameshift as a possible regulatory process [16]. FSscan presented in the study provides an algorithm to predict potential +1 frameshift products in *E. coli*.

#### 5.4 FSscan algorithm

FSscan is developed in Python (v2.4.3, Python Software Foundation, Hampton, NH) to search for potential +1 frameshift sites in the *E. coli* genome. The program assigns scores for a 16-nucleotide window along a gene sequence according to different effects of the stimulatory signals (S score) and interactions of the E-, P-, and A-site in the ribosome (E, P, and A scores, respectively) (Figure 5.1). A stimulatory signal in *E. coli* for +1 PRF can be a Shine Dalgarno (SD) - like sequence upstream of the frameshift site [17]. FSscan assigns zero to the S score if less than four base pairings can be formed between the six nucleotides upstream of the E-site position and the anti-SD sequence (3'UCCUCC5'); otherwise, FSscan assigns the number of base pairings divided by three to the S score (Eq.1).

$$\left. \begin{array}{l} (\text{Number of base pairings with UCCUCC}) < 4, S = 0 \\ (\text{Number of base pairings with UCCUCC}) \geq 4, S = (\text{Number of base pairings with UCCUCC})/3 \end{array} \right\} \text{Eq.1}$$

Sanders *et al.*, [18] suggested that zero frame codon:anticodon interactions in the E-site can affect frameshifting. The E score is calculated as  $\exp(-\Delta G_c)$ , where  $\Delta G_c$  is the codon:anticodon interaction [19] in the ribosome E-site. For the P-site, both zero

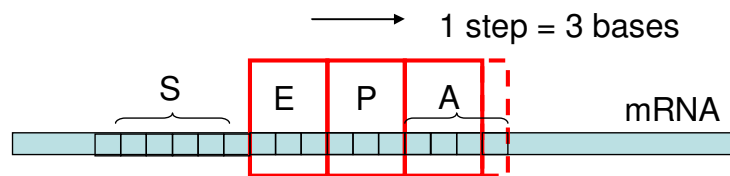


frame and +1 frame interactions can influence +1 frameshifting [20]. The P score in the program represents the stability difference between the zero frame and the +1 frame interactions for the P-site tRNA, normalized with the maximum stability difference obtained among 256 possible P-site sequences (Supplementary Data). The A score is the combination of the  $A_0$  score and the  $A_1$  score. The  $A_0$  score is the ratio of the arrival frequency, on the basis of transport by diffusion, of the near-cognate aa-tRNA versus the cognate aa-tRNA corresponding to the zero frame A-site codon [21], normalized with the maximum ratio of the arrival frequency obtained among 64 possible zero frame A-site codons. The  $A_1$  score is the ratio between the concentration of the cognate aa-tRNA for the +1 frame A-site codon to that of the cognate aa-tRNA for the zero frame A-site codon [21], normalized with the maximum concentration ratio obtained among 256 possible A-site sequences. For a stop codon in the zero frame A-site, the  $A_0$  and  $A_1$  scores were set to be 0.9 for TAG and TGA, and 0.6 for TAA. If the summation of the E, P, and A scores is less than three, the S score is then reset to zero (Eq. 2).

$$E + P + A < 3, S = 0, \text{ for any number of base pairings with UCCUCC} \quad \text{Eq.2}$$

Eq. 2 has a higher priority than Eq. 1, which means, as long as the summation of the E, P, and A score is less than three, the program assigns zero to the S score no matter how many base pairings can be formed between the mRNA sequence and the anti-SD sequence. The frameshift index (FSI) for a 16-nucleotide window is calculated as Eq. 3.

$$\text{Frameshift Index (FSI)} = S + E + P + A \quad \text{Eq.3}$$



$$FSI = S + E + P + A ; A = A_0 + A_1$$

S score is based on the number of base pairings with anti-Shine Dalgarno sequence.

E score is based on the tRNA:mRNA interaction in the E-site.

P score is based on the stability difference between the zero frame and the +1 frame interactions in the P-site.

A score is based on (1) the competition between the near-cognate aa-tRNA versus the cognate aa-tRNA for the zero frame A-site codon ( $A_0$  score) (2) the competition between the cognate aa-tRNA for the +1 frame A-site codon and the cognate aa-tRNA for the zero frame A-site codon ( $A_1$  score).

Figure 5.1. The scoring system for FSscan program. FSscan calculates scores for a 16-nucleotide window along the gene sequence. Each step is 3 nucleotides. FS index (FSI) =  $S + E + P + A$ .

A higher FSI suggests the sequence contains more features for +1 frameshifting. It is important to note that FSI is not set for quantitatively predicting the level of the +1 frameshifting, but rather how likely a sequence is a frameshift site.

## **5.5 Materials and methods**

### **5.5.1 Plasmids and bacterial strains**

*Escherichia coli* XL1 blue MRF<sup>+</sup> (Stratagene, La Jolla, CA) was used in all experimental studies. All constructs were verified by DNA sequencing. The construction of the dual fluorescence reporter was performed as described previously [14]. The control strain has both DsRed and enhanced green fluorescence protein (EGFP) coding sequences in frame. For the test strain, the linker sequences inserted between the two reporters contained predicted frameshift sequences followed by an in-frame stop codon and the downstream *egfp* in the +1 frame. The control strain expressed the DsRed-EGFP fusion protein from the reporter. The test strains expressed DsRed proteins as non-frameshift proteins (due to the stop codon in the linker sequence) and DsRed-EGFP fusion protein as frameshift proteins (because the stop codon is bypassed by +1 frameshifting). Table 5.1 lists the nucleotide sequences incorporated into the dual fluorescence reporter for testing +1 frameshift efficiency *in vivo* in this study. A negative control strain, *ran1*, was transformed with a plasmid containing a randomly designed linker (*rand*) inserted between the two fluorescence reporters with *egfp* in the +1 frame.

The first 915 nucleotides in *yehP* were PCR-amplified with the forward primer, *yehPf*, 5'-AAACTGCAGAATGTCTGAACTGAACGATCTTCTG -3' (PstI site underlined) and two reverse primers, *yehPr0* 5'-ATTGGTACCACGAGGATAATGACGCTTTTCGCTGG-3' and *yehPr1* 5'-ATTGGTACCCACGAGGATAATGACGCTTTTCGCTGG -3' (KpnI site

underlined) using *E. coli* genomic DNA as a template. The PstI/KpnI restricted PCR products were ligated with a PstI/KpnI-restricted pEGFP (Clontech, Mountain View, CA) vector to yield pYehP0 (using yehPr0 as the reverse primer for PCR) and pYehP1 (using yehPr1 as the reverse primer for PCR). The predicted frameshift sequence in pYehP1 was mutated by using QuikChange II site-directed mutagenesis kit (Stratagene) to create pYehPC. BsrGI/EcoRI restricted pYehP0, pYehP1, and pYehPC were ligated with a nucleotide sequence, 5'-GTACAAGCATCATCATCATCATTAAG-3', to create pYehP20, pYehP21, and pYehP2C to add a 6X-histidine tag downstream of *egfp*. KpnI/NcoI restricted pYehP20, pYehP21, and pYehP2C were ligated with a nucleotide sequence, 5'-CGTCTAGCTCTGGCTCTGGCTCTGGCAC-3', to create pYehP40, pYehP41, and pYehP4C to incorporate an in-frame stop codon and a flexible linker between *yehP* and *egfp*. *E. coli* strains transformed with pYehP40, pYehP41, and pYehP4C are named yehP40, yehP41, and yehP4C, respectively.

### 5.5.2 Fluorescence assay

Cells with the appropriate plasmids were cultured in 1 ml Luria-Bertani (LB) medium containing 100 µg/ml ampicillin in a 24-well plate for 24 hours at 37°C. The fluorescence was then measured by a plate reader (SpectraMax M5, Molecular Devices, Sunnyvale, CA). The fluorescence measurement was performed as described previously [14]. Frameshift efficiency (FS%) was obtained as the ratio of the green fluorescence to the red fluorescence for the test strains, normalized against the fluorescence ratio of the control strain. Statistical analysis was applied to all datasets according to Jacobs *et al.*, [22]. Eleven to twelve replicates for test strains and control strains were performed to satisfy the minimum sample requirement for statistical significance.

Table 5.1 Nucleotide sequences incorporated into the dual fluorescence reporter system for testing +1 frameshift efficiency *in vivo* in Chapter 5. *yehP*, *nuoE*, *pepP*, *cheA*, *ygchH*, and *yeal* are the top ranking candidates identified by FSscan. *glnD*, *yjgN*, and *cysD* are selected genes with one or two frameshifting features. *rand* is a randomly designed sequence to serve as a negative control.

Original Gene	16-nucleotide window with max FSI in the gene (the P-site position is underlined)						Strain (transformed with corresponding reporter plasmids)
<i>yehP</i>	GTG	GAG	TAT	<u>GGT</u>	CGG	C	yehP6
<i>nuoE</i>	GAG	CGG	TAT	<u>AAA</u>	TGA	A	nuoE6
<i>pepP</i>	AGT	GAG	ATA	<u>TCC</u>	CGG	C	pepP6
<i>cheA</i>	AGT	CGC	TAT	<u>CCC</u>	CGG	C	cheA6
<i>ygchH</i>	CCA	CTC	TAT	<u>TTT</u>	CGG	C	ygchH6
<i>yeal</i>	AAT	ATT	TAT	<u>AAT</u>	CGG	C	yeal6
<i>pspD</i>	CAG	CGT	TAT	<u>AAA</u>	AGG	T	pspD6
<i>glnD</i>	GGT	GGG	ATA	<u>AAA</u>	GCC	C	glnD6
<i>yjgN</i>	GAG	AGA	TAT	<u>TTT</u>	CTT	A	yjgN6
<i>cysD</i>	CAG	GGG	TAT	<u>TTT</u>	TAA	G	cysD6
<i>rand</i>	TCT	GGC	TCT	<u>GGC</u>	TGA	G	ran1
<i>yehP</i>	GTG	GAG	<b>TTA</b>	<u>GGT</u>	CGG	C	yehP7
(mutated sequence shown in bold)							

### **5.5.3 Western analysis**

Cells with the appropriate plasmids were cultured in 3 ml LB medium containing 100 µg/ml ampicillin in 17 ml round bottom tubes at 37°C. Aliquots of cells were harvested after 24hr cultivation and pelleted by centrifugation for 20 min at 4 °C and 4,000×g. The cell pellet was resuspended in 50 µl phosphate buffered saline per OD<sub>600</sub> and resolved by SDS-PAGE (10% w/v) using Tris-HCl. Immunoblot was performed as described by Gupta *et al.* [23], except rabbit anti-GFP (1:5000, Clontech) and alkaline phosphatase conjugated mouse anti-rabbit IgG antibody (1:10,000; Sigma, St. Louis, MO) were used as the primary and secondary antibodies, respectively.

### **5.5.4 Protein digestion**

yeh41 cell lysate was purified by Ni-NTA under denaturing conditions according to the manufacturer's protocol (Qiagen, Valencia, CA). The purified protein sample was exchanged into 0.2 M ammonium bicarbonate using Amicon Ultra 10 kDa molecular cutoff filter (Millipore, Billerica, MA). The buffer-exchanged sample was denatured and reduced by 6 M urea and 200 mM dithiothreitol (DTT) at room temperature for an hour. Then, the sample was alkylated by 200 mM iodoacetamide at room temperature for an hour in the dark. The remaining iodoacetamide in the sample was quenched by 200 mM DTT at room temperature for an hour and the sample was digested by trypsin (Promega, Madison, WI) at 37 °C for 14 hours. The digestion was stopped by decreasing the pH of the solution with 88% formic acid (FA) and vacuum dried, and the digested sample was reconstituted with 25 µL of 0.1% FA.

### **5.5.5 Liquid chromatography tandem mass spectrometry (LC-MS/MS)**

1.2 µL of the digested sample was separated by Dionex 3000 nLC system (Sunnyvale, CA) with an Acclaim PepMap 100 C18 trap column (300 µm × 5 mm, 5 µm, for the

on-line desalting at a flow rate of 30  $\mu$ L/min for 3 minutes) and an Acclaim PepMap 100 C18 analytical column (75  $\mu$ m  $\times$  15 cm, 3  $\mu$ m) at a flow rate of 250 nL/min. Peptides were eluted with gradients of 2-90% acetonitrile with 0.1% FA and the eluent was directly introduced into 4000 QTRAP MS through Nanospray II source (Applied Biosystems, Foster City, CA) for MRM study. To determine the appropriate MRM transitions that would be specific to the peptide of interest, the frameshift protein sequence was imported into the MIDAS Workflow software system (Applied Biosystems). The software generates a list of possible MRM transitions (Table 5.S2), including mass to charge ratios of precursor ions, fragment ions, and collision energy values for fragmentation. MS and MS/MS data obtained through MRM were searched within a custom sequence database that included the addition of the frameshift protein sequence. The spectral assignment of MS/MS were performed using ProteinPilot (v1.2 Applied Biosystems).

## **5.6 Results**

### **5.6.1 FSscan identifies a +1 frameshift hot spot in *prfB* gene**

FSscan successfully identifies the +1 frameshift site in *prfB*. Figure 5.2 shows the FSI along the *prfB* gene sequence. The FSI is at maximum when the ribosome P-site is positioned at the 25th codon in the coding sequence, the frameshift site for *prfB* in the literature [3].

### **5.6.2 Analysis of 4132 protein coding sequences in the *E. coli* genome reveals additional potential +1 frameshift candidates**

To identify potential +1 frameshifting sites, FSscan analyzed 4132 protein coding sequences in *E. coli* K12 MG1655 genome (Genbank: U00096). Because the FSI calculation requires an additional nucleotide downstream of the A-site codon, the 4132

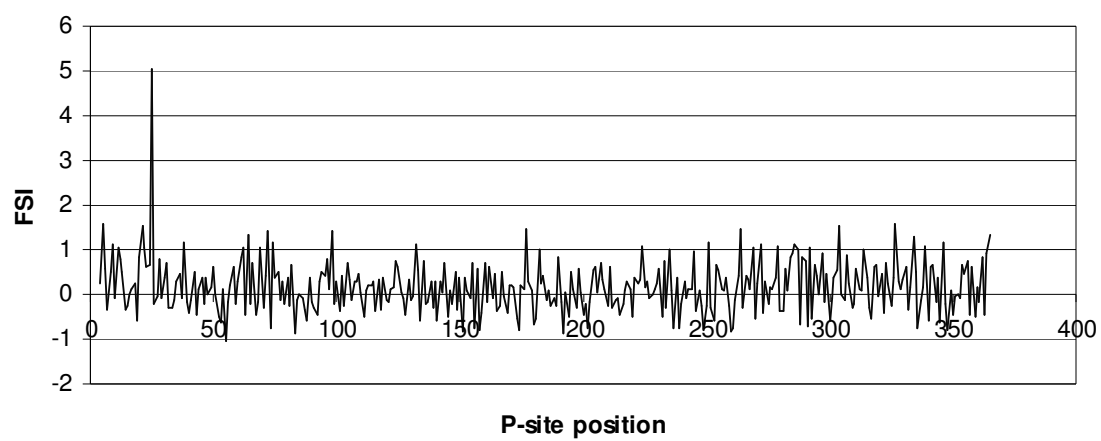


Figure 5.2. FSscan identifies the +1 frameshift site in *prfB*. A peak FSI is observed as the ribosome P-site is positioned at the 25<sup>th</sup> codon.



coding sequences were adjusted to include one more nucleotide downstream of the stop codon. The maximum FSI obtained in each protein coding sequence is plotted in Figure 5.3. *prfB*, whose expression has been shown to involve +1 PRF [3], has the highest FSI among all tested coding sequences (maximum FSI in *prfB* = 5.05). The next four highest ranking genes are *yehP*, *nuoE*, *pepP*, and *cheA*, with a maximum FSI 4.47, 4.39, 4.39, and 3.54 in their coding sequences, respectively. The potential +1 frameshift sequences in these genes are listed in Table 5.1. None of these candidates have been reported by previous approaches to identify +1 PRF genes [11, 12]. The other 4127 protein coding sequences all have a maximum FSI lower than 3.50.

### 5.6.3 *In vivo* examination of +1 frameshift sequences agrees with the program predictions

Several +1 frameshift candidates were examined *in vivo* by using a dual fluorescence reporter system. A randomly designed sequence with FSI=1.70 (*rand*, Table 5.1) was constructed to serve as a negative control strain (see Materials and Methods). Potential frameshift sequences from *yehP*, *nuoE*, *pepP*, and *cheA* resulted in FS% significantly higher than *rand* (Figure 5.4). A lower FS% was observed for sequences with FSI less than 3.5, suggesting that FSI 3.5 may serve as a threshold for identifying potential frameshift cassettes.

### 5.6.4 FSscan identifies *yehP* as a +1 frameshift candidate

*yehP* contains a potential +1 frameshift sequence with the second highest FSI, only after *prfB*. The predicted frameshifting sequence is GTG GAG **TAT** GGT CGG C (where each zero frame codon is separated by a space and the P-site position for obtaining the maximum FSI is underlined). In this sequence, an ATG in the +1 frame (shown in bold in the sequence above) together with an upstream GGAG may result in

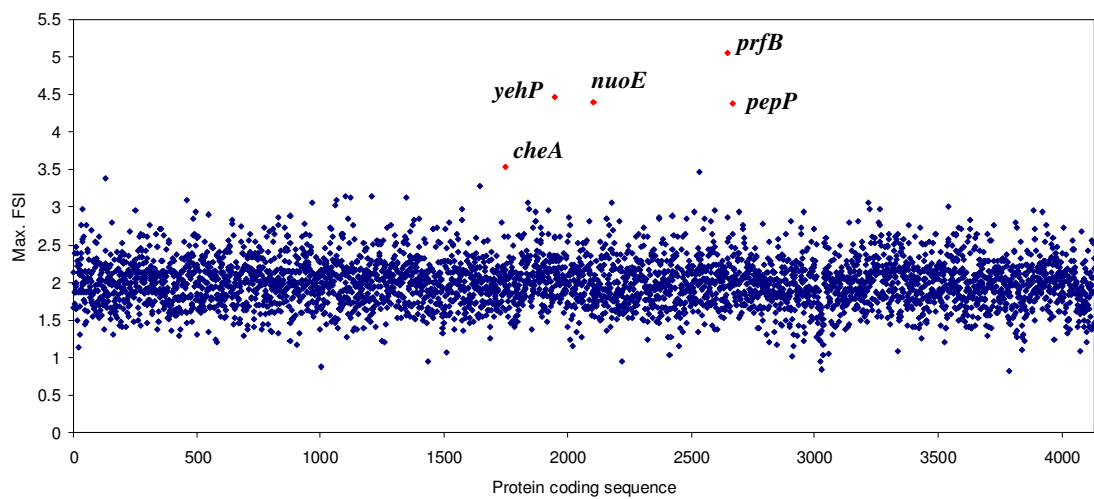


Figure 5.3. Maximum FSI in each of the 4132 *E. coli* protein coding sequences. Five genes with a maximum FSI above 3.5 are indicated in red. *prfB* has the maximum FSI 5.05. *yehP* has the maximum FSI 4.47. *nuoE* has the maximum FSI 4.39. *pepP* has the maximum FSI 4.39. *cheA* has the maximum FSI 3.55.

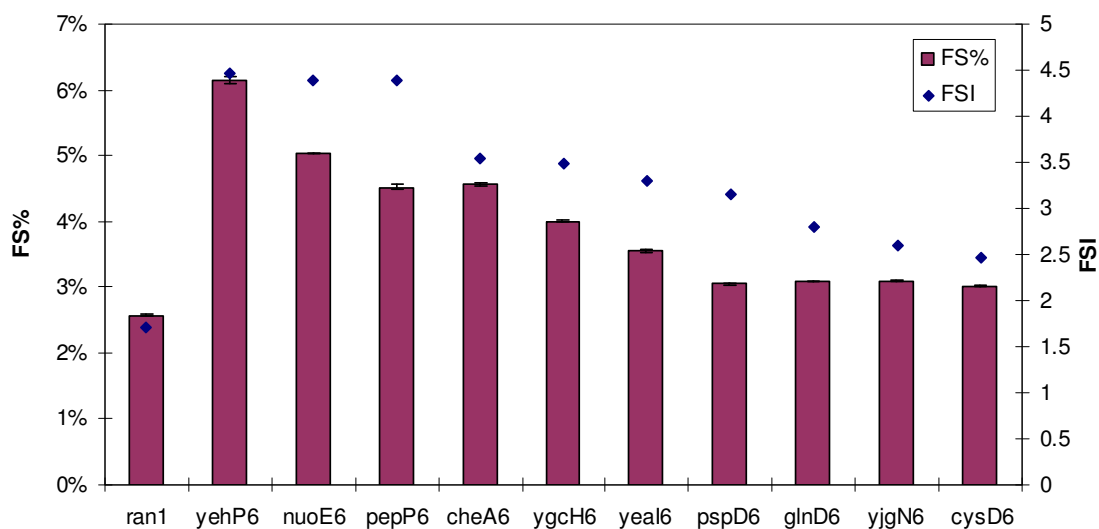


Figure 5.4 Frameshift efficiency (FS%) for potential frameshift sequences identified by FSscan. The histogram indicates the experimentally observed FS% for different test strains listed in Table 5.1. Error bars show the standard deviation. Diamonds demonstrate the program calculated FSI for the potential frameshift cassettes (sequences are shown in Table 5.1).

internal translation, causing non-frameshifting based EGFP expression in the dual reporter system. To further confirm *yehP* as a candidate +1 PRF gene, the sequence was mutated to GTG GAG **TTA** **GGT** CCG C (mutation shown in bold) to remove ATG in the +1 frame while keeping a weaker E-site interaction (yehP7 in Table 5.1). A small decrease in FS% was observed (Figure 5.5), but the mutation still resulted in a significantly higher FS% as compared to the negative control strain, *ran1* (Figure 5.4). This observation suggests that the higher FS% for yehP6 is not likely due to the internal translation of EGFP starting from the linker sequence.

To study the frameshift site in *yehP*, the fusion constructs yehP40, yehP41, and yehP4C were made with *egfp* 3' to *yehP* (Figure 5.6a). Proteins from cell lysate were subjected to Western analysis. Protein bands with molecular weight 63 kDa, the expected mass for the fusion protein, were observed for yehP40 and yehP41. Interestingly, no or very few proteins with this mass were observed when the potential frameshift sequence was mutated to GTG GAG **TCT** **TGT** CGA C to remove frameshifting features (yehP4C, mutated nucleotides shown in bold) (Figure 5.6a and 6b). The result suggests that the +1 frameshift event is specific to the predicted sequence.

Proteins from yehP41 cell lysate were purified, buffer-exchanged and digested by trypsin. The digest was analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) using multiple reaction monitoring (MRM). MRM is a highly sensitive scanning technique for peptide identification. The greater specificity is achieved by fragmenting the analyte and monitoring both parent and one or more product ions simultaneously (see review by Kitteringham *et al.*, [24]). Figure 5.7 presents the amino acid sequence derived from the frameshift site and the tryptic peptides observed by MRM. The presence of the peptide VQLGGGTNIASAVEYGGNLLNNQR (Figure

5.S3 in the Supplementary Data), whose coding sequence spans the potential frameshift site, is a result of the +1 frameshifting at the 291<sup>st</sup> codon, GTT CGG C (where the P-site position is underlined), in *yehP*. This result further confirms the frameshift site in *yehP*, as suggested by FSscan.

For +1 frameshifting at the 291<sup>st</sup> codon in *yehP*, the ribosome encounters a stop codon 15 codons downstream of the frameshift site. As a result, the frameshift product is 303 amino acids in length, which is 75 amino acids shorter than the non-frameshift *yehP* product. Importantly, *yehP* is highly conserved in different *E. coli* strains and is also observed in several other eubacteria (Table 5.2). The consensus of the *yehP* frameshift cassette for the 31 sequences in Table 5.2 is shown by a sequence logo (Figure 5.8) [25, 26]. Only a minor diversity is observed at position 1, 6, 12, and 14 in the 16-nucleotide frameshifting window.

## **5.7 Discussion**

### **5.7.1 The scoring system**

In FSscan, the S score represents the stimulatory effect on +1 frameshifting. FSscan assigns zero to the S score for less than four base pairings between the six nucleotides upstream of the E-site and the anti-SD sequence (Eq.1). Eq.1 implies that at least four base pairing between mRNA and the anti-SD sequence are required to reveal the stimulatory effect. FSscan identifies *yehP* as the second best candidate for +1 frameshifting by using four as a threshold value in Eq.1, while the program identifies *cheA* as the second best candidate by using five as a threshold value. The *in vivo* observation that *yehP*6 results in higher frameshift efficiency than *cheA*6 (Figure 5.4) suggests that four base pairings could be sufficient to induce a stimulatory effect. In addition, FSscan assigns zero to the S score if the summation of the E, P, and A scores

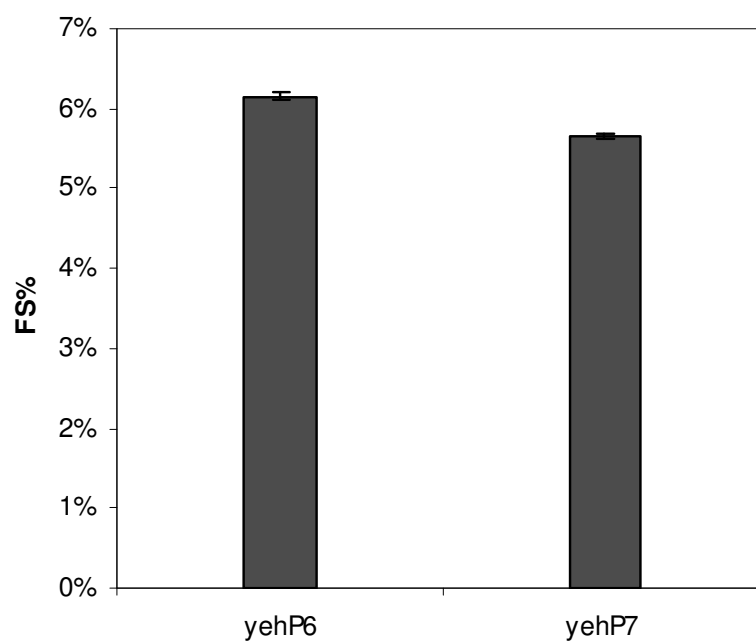
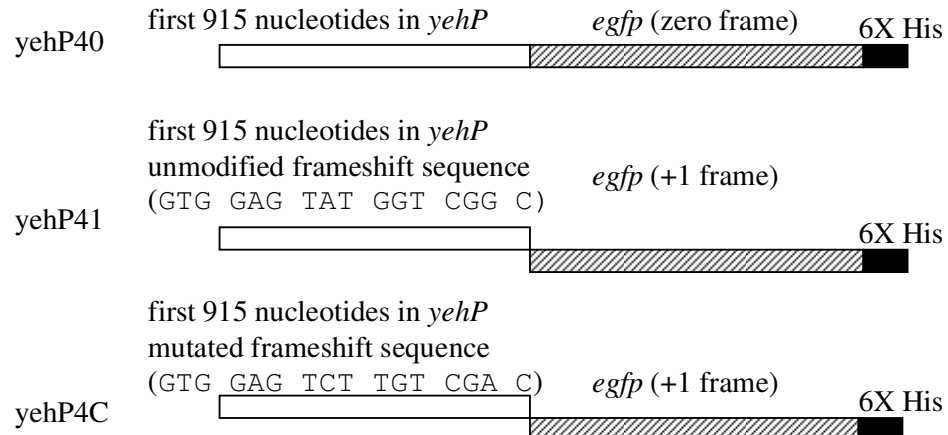


Figure 5.5. Frameshift efficiency (FS%) for yehP6 and yehP7. In yehP6, the linker inserted between the two fluorescence reporters contains the predicted *yehP* frameshift sequence: GTG GAG TAT GGT CGG C. In yehP7, the frameshift sequence is mutated to GTG GAG TTA GGT CGG C (where zero frame codons are separated by spaces).

(a)



(b)

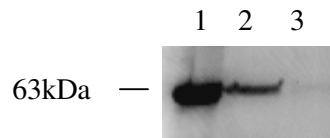
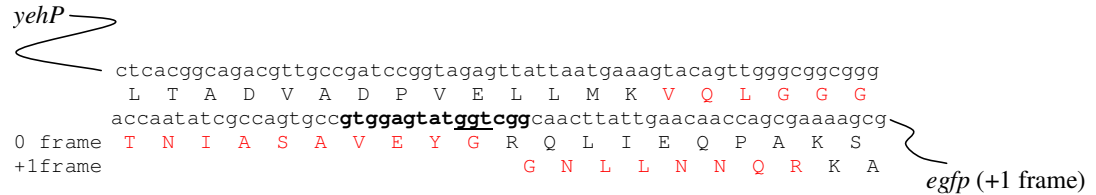


Figure 5.6. (a) The nucleotide sequence design for yehP40, yehP41, and yehP4C. (b) Western blot for the cell lysate to detect the frameshift protein. Lane 1: total lysate from yehP40. Lane 2: total lysate from yehP41. Lane 3: total lysate from yehP4C. The amount of the protein loaded for yehP40 is one third of the amount of the protein for yehP41 and yehP4C.

(a)



(b)

```

MSELNDLLTTRELQRWRLILGEAAETTLCLDDNARQIDHALEWLYGRDPERLQRGERSG
GLGGSNLTTPEWINSIHTLFPQQVIERLESDAVLRYGIEDVVTNLDVLERMQPSESLLRA
VLHTKHLMNPEVLAAARRIVCQVVEIMARLAKEVRQAFSGVRDRRRRSFIPLARNFDFK
STLRANLQHWHPQHGLYIESPRFNSRIKRQSEQWQLVLLVDQSGSMVDSVIHSAVMAAC
LWQLPGIRTHLVAFDTSVVDLTADVADPVELLMKVQLGGGTNIASAVEYGNLLNNQRKA
SLSSWVPSSSGSGSGTMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGLLT
LKFICTTGKLPVPWPTLVTTLTYGVCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDD
GNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNNSHNVYIMADKQKNGIKV
NFKIRHNIEDGSGVQLADHYQQNTPIGDPVLLPDNHYLSTQSALS KDPNEKRDHMLLEF
VTAAGITLGMDELYKHHHHHHH-

```

Red letters: peptides identified with confidence level higher than 95% from ProteinPilot

Yellow background: peptides originated from YehP

Black-outlined box: peptides spanning the frameshift site

Blue background: peptides originated from EGFP

Figure 5.7. Nucleotide and amino acid sequence for the YehP-EGFP frameshift protein in yehP41. (a) The nucleotide and amino acid sequence for the predicted frameshift region in YehP-EGFP. The predicted frameshift sequence is shown in bold, with the P-site codon underlined. The zero frame and the +1 frame amino acid sequences are shown under the nucleotide sequence. The peptide spanning the frameshift site, with the zero frame translation before the site and the +1 frame translation after the site, is shown in red. (b) Amino acid sequence for the frameshift protein in yehP41 strain. The YehP-EGFP was expressed as a result of +1 frameshifting. Tryptic peptides observed by MRM are marked in red (>95% confidence level). The sequence coverage is 21.7%.



is less than three (Eq.2). Eq.2 implies that for a less prominent synergic effect of the E-, P-, and A-site for +1 frameshifting, the stimulatory effect by SD: antiSD interaction is negligible.

The E score in the program represents the effect of E-site interaction on +1 frameshifting. FSscan calculates the E score as  $\exp(-\Delta G_c)$ , where  $\Delta G_c$  is the codon:anticodon interaction [19] in the ribosome E-site. The interaction in ribosome E-site has been shown to affect the reading frame maintenance [14, 18, 27-30]. Weaker codon:anticodon interactions in the ribosome E-site have also been observed to result in a higher +1 frameshift efficiency [14,18]. Notably, FSscan does not account for different tRNA:ribosome interactions in the E-site. While the tRNA:ribosome interactions are important for the E-site interaction, there has not been a well-established method to estimate these interactions. Previously, it has been suggested that a major fraction of the E-site tRNA binding is contributed by the binding of the 3'-terminal adenine to the ribosome [31]. Because the 3'-terminal adenine is conserved in all *E. coli* tRNAs, FSscan assumes a similar level of tRNA:ribosome interactions for different tRNAs and considers only codon:anticodon interactions in the E-site.

The P score represents for the stability difference between the +1 frame and the zero frame interaction for the P-site tRNA. FSscan assumes the stability difference between the +1 frame and the zero frame interaction ( $\Delta$  stability\*) as  $M_1S_1 - M_2S_0$ , where  $S_1$  is the stability of the +1 frame interaction,  $S_0$  is the stability of the zero frame interaction, and  $M_1$  and  $M_2$  are weighting factors. A separate data fitting program suggests  $M_1$  and  $M_2$  as 0.63 and 0.26, respectively, for the best linear correlation between the  $\Delta$  stability\* and the logarithm of +1 frameshift efficiency observed by

Table 5.2 BLAST result for *yehP*. blastn was used as the algorithm to search the nucleotide collection database in National Center for Biotechnology Information's website. The search was optimized for highly similar sequences.

Accession	Description	Max score	Total score	Query coverage	E value	Max ident <sup>a</sup>
CP000948.1	<i>Escherichia coli</i> str. K12 substr. DH10B, complete genome	2254	2290	100%	0.0	100%
AP009048.1	<i>Escherichia coli</i> str. K12 substr. W3110 DNA, complete genome	2254	2290	100%	0.0	100%
U00096.2	<i>Escherichia coli</i> str. K-12 substr. MG1655, complete genome	2254	2290	100%	0.0	100%
U00007.1	47 to 48 centisome region of <i>E.coli</i> K12 BHB2600	2254	2254	100%	0.0	100%
CU928160.2	<i>Escherichia coli</i> str. IAI1 chromosome, complete genome	2119	2155	100%	0.0	100%
AP009240.1	<i>Escherichia coli</i> SE11 DNA, complete genome	2095	2132	100%	0.0	100%
CP000800.1	<i>Escherichia coli</i> E24377A, complete genome	2095	2132	100%	0.0	100%
CP000036.1	<i>Shigella boydii</i> Sb227, complete genome	2095	2168	100%	0.0	100%
AB426057.1	<i>Escherichia coli</i> O111:H- DNA, genomic island GEI2.21	2087	2087	100%	0.0	98%
CP000034.1	<i>Shigella dysenteriae</i> Sd197, complete genome	2087	2160	100%	0.0	100%
CP000946.1	<i>Escherichia coli</i> ATCC 8739, complete genome	2056	2092	100%	0.0	100%
CP000802.1	<i>Escherichia coli</i> HS, complete genome	2032	2068	100%	0.0	100%
AE005674.1	<i>Shigella flexneri</i> 2a str. 301, complete genome	1992	2065	100%	0.0	100%
AE014073.1	<i>Shigella flexneri</i> 2a str. 2457T, complete genome	1992	2065	100%	0.0	100%
AE014075.1	<i>Escherichia coli</i> CFT073, complete genome	1976	2085	100%	0.0	100%
CU928164.2	<i>Escherichia coli</i> str. IAI39 chromosome, complete genome	1961	2033	100%	0.0	100%
BA000007.2	<i>Escherichia coli</i> O157:H7 str. Sakai DNA, complete genome	1961	2033	100%	0.0	100%
AE005174.2	<i>Escherichia coli</i> O157:H7 EDL933, complete genome	1961	2033	100%	0.0	100%
CP001164.1	<i>Escherichia coli</i> O157:H7 str. EC4115, complete genome	1953	2025	100%	0.0	100%
CP000970.1	<i>Escherichia coli</i> SMS-3-5, complete genome	1937	2009	100%	0.0	100%
CU928162.2	<i>Escherichia coli</i> str. ED1a chromosome, complete genome	1913	2021	100%	0.0	100%
FM180568.1	<i>Escherichia coli</i> O127:H6 E2348/69 complete genome, strain E2348/69	1905	1977	100%	0.0	100%
CU928161.2	<i>Escherichia coli</i> str. S88 chromosome, complete genome	1897	2006	100%	0.0	100%
CP000468.1	<i>Escherichia coli</i> APEC O1, complete genome	1897	2006	100%	0.0	100%
CP000243.1	<i>Escherichia coli</i> UTI89, complete genome	1897	2006	100%	0.0	100%
CU928158.2	<i>Escherichia fergusonii</i> str. ATCC 35469T chromosome, complete genome	1850	1924	100%	0.0	95%
CP000247.1	<i>Escherichia coli</i> 536, complete genome	1850	1958	100%	0.0	100%
CU928163.2	<i>Escherichia coli</i> str. UMN026 chromosome, complete genome	1842	1914	100%	0.0	100%
CU651637.1	<i>Escherichia coli</i> LF82 chromosome, complete sequence	1818	1926	100%	0.0	100%
AP000400.1	Enterobacteria phage VT1-Sakai genomic DNA, prophage inserted region in <i>Escherichia coli</i> O157:H7	1542	1542	81%	0.0	96%
CP000038.1	<i>Shigella sonnei</i> Ss046, complete genome	603	675	29%	8e-169	100%

<sup>a</sup> maximum identities



Figure 5.8. Sequence conservation of the predicted frameshift cassette in *yehP*. The sequence logo was generated by aligning 31 sequences in Table 5.2.

Curran (1993) [20] (Supplementary Data). The weighting factor for the +1 frame stability is 2.4-fold larger than that for the zero frame stability. Interestingly, zero frame duplexes are in general cognate but the realigned complexes contain a much wilder array of pairing and stabilities. Taken together, a favorable +1 frame interaction in the P-site may contribute more than an unTable 5.zero frame interaction to a higher +1 frameshift efficiency.

FSscan accounts for two A-site features that enhance +1 frameshifting: (1) the competition between the cognate and the near cognate aa-tRNA for the zero frame A site codon ( $A_0$  score); (2) the competition between the cognate aa-tRNA for the zero frame A-site codon and the cognate aa-tRNA for the +1 frame A-site codon ( $A_1$  score). A ribosome pause because of a stop codon or a rare codon in the A-site is a key factor for +1 frameshifting [32, 33]. It has been shown that the competition between the near-cognate aa-tRNA and the cognate aa-tRNA to the ribosome A-site plays an important role on the translation rate [21]. The imbalance of the zero frame A-site tRNA and the +1 frame A-site tRNA was also shown to enhance +1 frameshifting [34]. Three +1 frameshift candidates, *yehP*, *pepP* and *cheA*, all have C G G C in the A-site (where the zero frame codon is separated by the space). While the average A score is 0.44, the A score for C G G C is 1.58. C G G has one cognate tRNA,  $tRNA^{Arg}_{CCG}$ , with 639 molecules per cell, and four near-cognate tRNAs,  $tRNA^{Arg}_{ACG}$ ,  $tRNA^{Gln}_{CUG}$ ,  $tRNA^{Leu}_{CAG}$ , and  $tRNA^{Pro}_{CGG}$ , with 4752, 881, 4470, and 900 molecules per cell, respectively [21]. The fact that near-cognate tRNAs outnumber cognate tRNAs for C G G results in a competition between these tRNAs for the ribosome A-site. In addition, the concentration of the cognate tRNA for the +1 frame A-site codon (G G C) is about seven-fold higher than that for the zero frame A-site codon (C G G). These two features may result in a longer pause during translation, making C G G C a likely A-site codon

for +1 frameshifting. The other candidate *nuoE* has TGA A in the A-site. The A score for TGA A is 1.8, which is also much higher than the average A score.

FSI for a 16-nucleotide window sums up S, E, P, and A scores. The S score ranges from 0 to 2. The E score ranges from 0 to 1. The P score ranges from -1 to 1. The A score ranges from 0 to 2 because it combines A<sub>0</sub> and A<sub>1</sub>, each ranging from 0 to 1. As a result, FSscan weighs the stimulatory, P-site, and A-site effects more than the E-site effect. This algorithm is supported by the kinetic model of +1 PRF, which suggested that +1 frameshift efficiency is more sensitive to the change in the stimulatory signal, P-site, and A-site effects [14].

### 5.7.2 Analysis of six reading frames and pseudogenes

Analysis of the six reading frames of the *E. coli* genome by FSscan reveals that 192 sequences have FSI higher than 3.5. 83 of these sequences are located in the annotated coding regions, but only five sequences are in-frame with the start codon. The five cassettes are in *prfB*, *yehP*, *nuoE*, *pepP*, and *cheA*. This result is consistent with the analysis of the 4132 protein coding sequences (Figure 5.3). The function of intergenic sequence with FSI higher than 3.5 is not clear and requires further investigation. In addition, none of the 163 pseudogenes in the *E. coli* genome had a maximum FSI higher than 3.5 (data not shown).

### 5.7.3 *yehP*

*yehP* contains a potential +1 frameshift site with the second highest FSI, only after *prfB*. The predicted frameshift site in *yehP* is highly conserved in different *E. coli* strains (Table 5.2 and Figure 5.8). The potential cassette, GTG GAG TAT GGT CGG C (the zero frame is separated by a space and the P-site position is underlined), forms

four base pairings with the anti-SD sequence and allows a weaker interaction in the E-site. In the P-site, tRNA<sup>Gly</sup><sub>GCC</sub> may form two canonical base pairings with the +1 frame although a central position mismatch can also occur. Notably, it has been proposed that less than two base pairings in the shifted codon:anticodon complex may be sufficient for the efficient frameshifting [35]. In a more extreme case, mRNA sites with little or no potential for canonical base pairing with the peptidyl-tRNA in the ribosome can also be used as landing positions for ribosomal bypassing [36]. In the A-site, CCG is one of the four codons with the highest near-cognate tRNA competition [21]. All of these features make *yehP* a potential +1 frameshifting candidate.

To date, the function of the *yehP* product is not well described in the literature. A known +1 PRF case in *E. coli* is the expression of RF2 from *prfB* gene [3]. RF2 frameshifting is auto-regulated, meaning higher frameshift efficiency is driven by a lower level of the frameshift products [3]. It is suggested that this auto-regulation property may be evolved to evade a newly discovered fidelity control system: the ribosome would trigger a pre-mature termination of protein synthesis when a mismatch P-site interaction is presented [37]. RF2 frameshifting occurs more frequently when RF2 level is low, making it more difficult for ribosomes to trigger early termination in the presence of mismatch P-site. Whether *yehP* has involved in any regulation feedback loop or other mechanisms to escape from this fidelity control mechanism is uncertain. A *yehP* knockout *E. coli* strain was previously shown to result in a different swarming phenotype [38]. *yehP* was suggested to have been introduced to the *E. coli* genome by the horizontal gene transfer [39]. The predicted frameshift product is 75 amino acids shorter than the standard decoding product. The function of the *yehP* frameshift protein remains unclear and needs to be investigated further.

#### 5.7.4 Other frameshift-prone sequences

FSscan did not identify several shift-prone sequences observed experimentally in previous studies [40, 41]. *argI* was found to have a high level of +1 frameshifting at the very beginning of the coding sequence, UUU UAU [40]. However, the maximum FSI in the gene is relatively low (2.0 for the P-site at the 110<sup>th</sup> codon). For the P-site positioned at the fourth codon UUU, FSI equals 0.38. Because *argI* frameshifting does not involve ribosomal pausing at a stop codon or a hungry codon in the A-site, the recoding may be achieved through mechanisms not considered by FSscan. In addition, CCC TGA containing genes, *pheL*, *yjeF*, *ykgD* and *yrhB*, were also shown to result in a higher level of +1 frameshifting [41]. Notably, these sequences do not form more than three base pairings with the anti-SD sequence and their E-site interactions are relatively strong, which result in lower FSI. It is possible that a slippery sequence in the P-site (i.e. P-site tRNA can form complementary interactions with the +1 frame) along with a stop codon in the A-site can efficiently induce +1 frameshifting, which FSscan does not consider. On the other hand, not all of the CCC TGA containing genes promotes efficient +1 frameshifting, suggesting different mechanisms may be involved for *pheL*, *yjeF*, *ykgD* and *yrhB* frameshifting. As growing numbers of the +1 frameshifting features are discovered, these features can be incorporated into FSscan to better predict frameshift sites.

#### 5.7.5 FSscan as a bioinformatic program to search for novel +1 frameshift sequences

FSscan locates a 16-nucleotide sequence with features for stimulatory signals, E-, P-, and A-site effects in the *E. coli* genome. As compared to previous +1 frameshift site searching programs [11, 12], FSscan differs in several major ways: (1) FSscan is not limited to a specific P- or A-site codon. Instead, FSscan looks for any P-site codon

with a higher opportunity for tRNA rearrangement and any A-site codon with a higher possibility for a ribosome pausing during translation; (2) The algorithm does not search for overlapping genes. Thus it is not necessary that predicted frameshifting cassettes yield C-terminally extended fusion products; (3) FSscan is intended for searching the *E. coli* genome, because the tRNA data for the score calculation and the experimental system are specific to *E. coli*. FSscan may be directly applied to screen the genome of *E. coli* bacteriophage, whose proteins can be translated by using *E. coli* ribosomes and tRNA pool. The strategy can be extended to other organisms with minor adjustments for the scoring system. (4) FSscan predicts how likely a sequence is a frameshift site, but not the +1 frameshift efficiency. (5) FSscan needs no prior knowledge of the mRNA secondary structure involved in recoding. This method can be modified by varying the size of the recoding window to include mRNA structures serving as stimulatory signals.

### 5.8 Supplementary data

The estimation of the P score

The stability of the zero frame and the +1 frame codon: anticodon interactions in the P-site were estimated according to Curran's study (1993) (Table 5.S1). Fig. S1a shows a good correlation between log frameshift efficiency (FS%) and the stability difference ( $\Delta$  stability), which is the stability of the +1 frame interaction ( $S_1$ ) minus that of the zero frame interaction ( $S_0$ ). To improve the correlation between  $\log(\text{FS}\%)$  and  $\Delta$  stability, we assumed  $\Delta \text{ stability}^* = M_1 S_1 - M_2 S_0$ , where  $M_1$  and  $M_2$  are weighting factors for  $S_1$  and  $S_0$ , respectively. The sum of squared errors is calculated as  $SS_{err}^2 = \sum_i (\log(\text{FS}\%)_i - \Delta \text{ stability}^*_i)^2$ , where  $i$  refers to different codons ( $i = 1-32$ , for 32 codons in Curran's study, 1993). Matlab (v.7.6.0, The MathWorks, Inc., Natick, MA) was used to obtain the values of  $M_1$  and  $M_2$  that resulted in the minimum



sum of the squared errors. The program identified a pair value of  $M_1$  and  $M_2$  as 0.63 and 0.26, respectively. The  $R^2$  value increased from 0.56 to 0.63 when  $M_1$  and  $M_2$  were applied. (Fig. S1a and S1b). Consequently, the P score is calculated as  $(0.63S_1 - 0.26S_0)/\text{maximum}(0.63S_1 - 0.26S_0)$  among 256 P-site codons. For example, for a UUC U (where the zero frame codon is separated by the space) in the P-site (anticodon 3'AAG5'),  $S_0 = 2 + 2 + 3 = 7$ ,  $S_1 = 2 + (-1) + 1 = 2$ ,  $0.63S_1 - 0.26S_0 = (-0.56)$  and the P score =  $(-0.56)/3.33 = -0.1682$ , where 3.33 is the  $0.63S_1 - 0.26S_0$  value for CCC C, which is the maximum value obtained among the 256 P-site codons.

In addition, UUU U, AAA A, and AAA U in the P-site (where the zero frame codon is separated by the space) showed lower frameshift efficiency *in vivo* (Fig.S2). As the result, for these three sequences in the P-site, the P-site score was set to 0.3.

Table 5.S1 The stability of the base pairing in the ribosome P-site according to Curran [20]. Marked in grey are assumed values in this study. A, G, C and U are the standard bases; I = inosine; V = uridine-5-oxyacetic acid; Q = queuosine; S = 5-methylaminomethyl-2-thiouridine; E = unidentified derivative of U that pairs with A; F = modified pyrimidine that pairs with A. 1st, 2nd, and 3rd mean the position in the codon.

	1st	2nd	3rd
AU	2	2	2
AC	0	-1	0
AG	0	-2	0
AA	0	-2	0

	1st	2nd	3rd
CU	0	0	0
CC	-1	-1	-1
CG	3	3	3
CA	0	-1	0

	1st	2nd	3rd
GU	1	0	1
GC	3	3	3
GG	0	-2	0
GA	0	-2	0

	1st	2nd	3rd
UU	0	1	0
UC	0	0	0
UG	1	0	1
UA	2	2	2

	1st	2nd	3rd
IU	1	0	1
IC	1	0	1
IG	0	-2	0
IA	1	0	1

	1st	2nd	3rd
VU	1	0	1
VC	0	0	0
VG	1	0	1
VA	1	0	1

	1st	2nd	3rd
QU	1	0	1
QC	1	0	1
QG	0	-2	0
QA	0	-2	0

	1st	2nd	3rd
SU	1	0	1
SC	0	0	0
SG	1	0	1
SA	1	0	1

E	1st	2nd	3rd
EU	0	0	0
EC	0	0	0
EG	1	0	1
EA	1	0	1

F	1st	2nd	3rd
FU	0	0	0
FC	0	0	0
FG	0	0	0
FA	1	0	1

Table 5.S2 Peptides detected by MRM and identified by ProteinPilot

Origin	Conf <sup>a</sup>	Sequence	Modifications	Exp MW <sup>b</sup>	Theor MW <sup>c</sup>	dMass <sup>d</sup>	Prec m/z <sup>e</sup>	Frag m/z	z <sup>g</sup>	CE <sup>h</sup>
yehP	99	LILGEAAETTLCLDDNAR	Carbamidomethyl(C)	2030.981	2030.994	-0.01317	678.001	760.354	3	32.832
yehP	99	YGIEDVVTNLDVLER		1733.87	1733.884	-0.01344	578.964	631.336	3	28.474
yehP	99	HLMNPEVLAAAR		1320.683	1320.697	-0.01411	441.235	501.310	3	22.414
yehP	99	RIVCQVVEEIMAR	Carbamidomethyl(C)	1601.825	1601.838	-0.01307	534.949	619.319	3	26.538
yehP	99	IVCQVVEEIMAR	Carbamidomethyl(C)	1445.727	1445.737	-0.00967	723.871	847.430	2	41.194
linker	99	VQLGGGTNIASAVEYGGNLLNNQR		2444.225	2444.241	-0.01556	815.749	871.470	3	38.893
EGFP	99	FSVSGEGEGDATY GK		1502.643	1502.653	-0.00911	752.329	840.369	2	42.616
EGFP	98	SAMPEGYVQER		1265.562	1265.571	-0.00954	633.788	751.369	2	36.689
EGFP	99	AEVKFEGDTLVNR		1476.743	1476.757	-0.01407	493.255	501.310	3	24.703

<sup>a</sup> Confidence level (%)

<sup>b</sup> Experimental molecular weight (Da)

<sup>c</sup> Theoretical molecular weight (Da)

<sup>d</sup> Delta mass (Da) = Exp MW – Theor MW

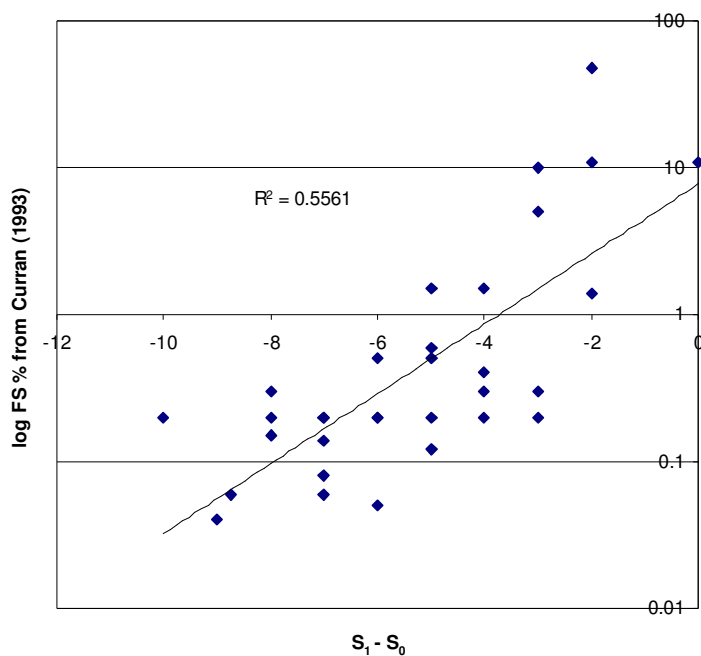
<sup>e</sup> Precursor ion *m/z* in a MRM transition

<sup>f</sup> Fragment ion *m/z* in a MRM transition

<sup>g</sup> Charge state of a precursor ion in a MRM transition

<sup>h</sup> Collision energy (V) used for a MRM transition

(a)



(b)

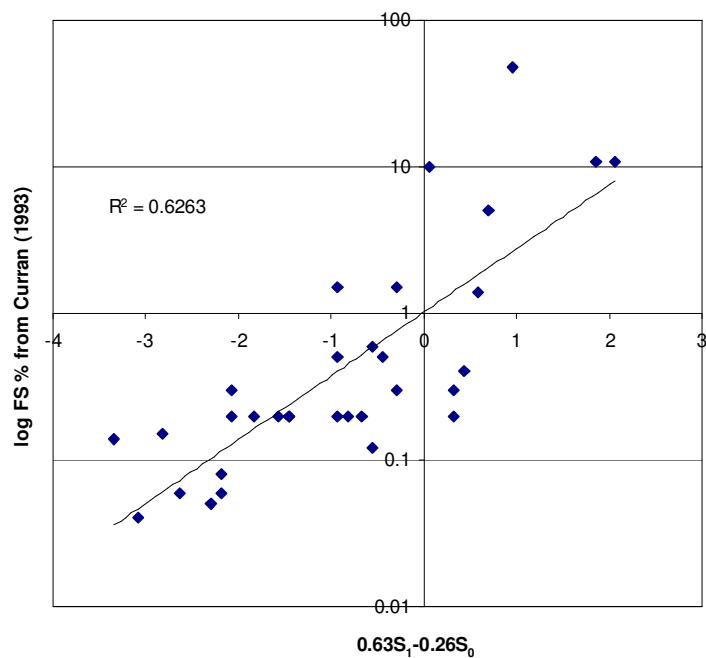
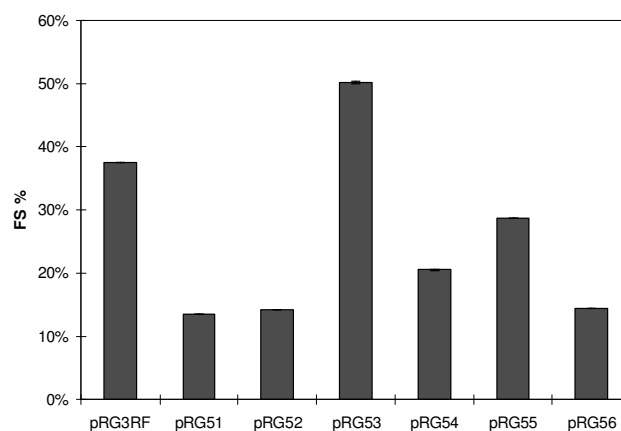


Figure 5.S1. The correlation between log frameshift efficiency (FS%) and the stability difference in the P-site. (a) The stability difference is estimated as  $S_1 - S_0$  (where  $S_1$  is the stability of the +1 frame interaction and  $S_0$  is the stability of the zero frame interaction in the P-site). (b) The stability difference is estimated as  $0.63S_1 - 0.26S_0$ .



Plasmid	Linker sequence						
pRG3RF	agg	ggg	tat	<b>ctt</b>	tga	cta	
pRG51	agg	ggg	tat	<b>ttt</b>	tga	cta	
pRG52	agg	ggg	tat	<b>aaa</b>	tga	cta	
pRG53	agg	ggg	tat	<b>ccc</b>	tga	cta	
pRG54	agg	ggg	tat	<b>gtt</b>	tga	cta	
pRG55	agg	ggg	tat	<b>cct</b>	tga	cta	
pRG56	agg	ggg	tat	<b>aaa</b>	agg	tga	

Figure 5.S2. Frameshift efficiency (FS%) for selected P-site codons. Error bars indicate the standard deviation. The linker sequences incorporated into the dual fluorescence reporter system are shown in the right, where the P-site codons in the frameshift site are shown in bold.

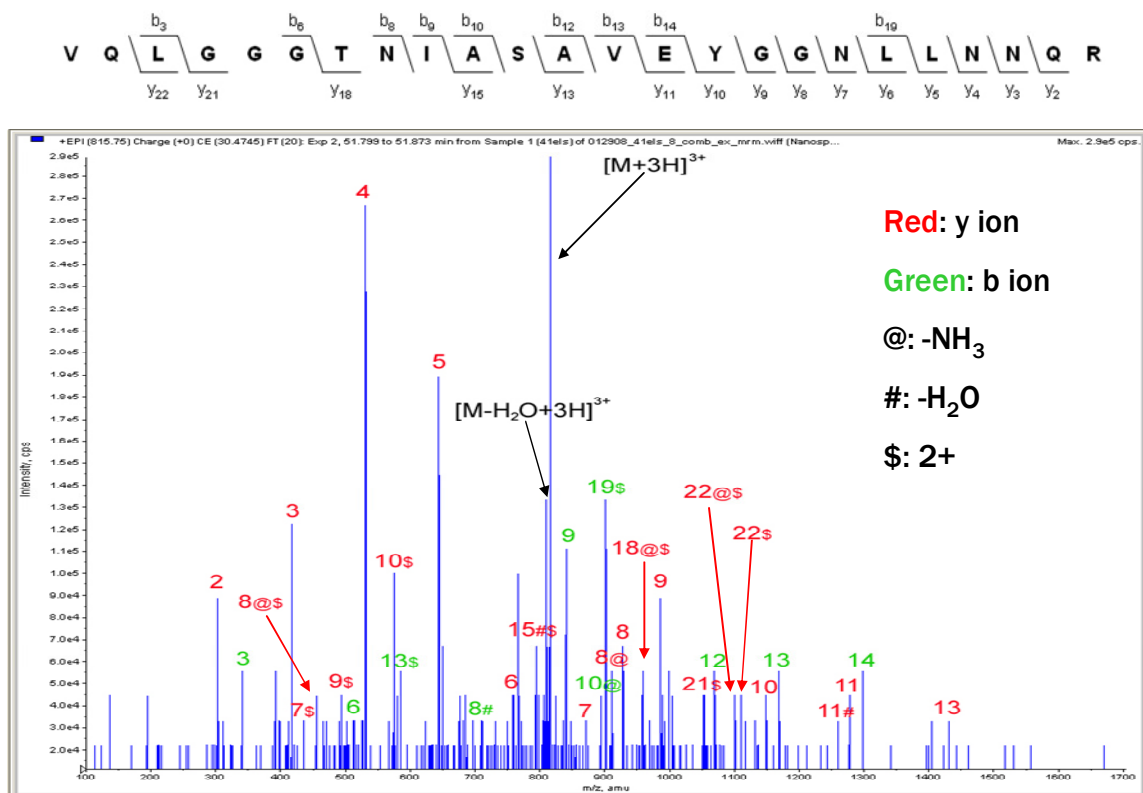


Figure 5.S3. Tandem mass spectrum (MS/MS) of the peptide derived from the predicted frameshift site in *yehP*.

### **5.9 Conclusion**

FSscan performs a mechanistic-based genetic algorithm search for potential +1 frameshift sites in *E. coli*. The program successfully identifies *prfB* as a +1 frameshift candidate and predicts the frameshift site in this gene. Other predicted frameshift cassettes are shown to result in frameshift efficiency higher than a randomly designed sequence *in vivo*. These results suggest that the synergistic effects of ribosome E-, P-, and A-sites are functionally important for +1 frameshifting. Importantly, FSscan provides the ability to perform a genome-wide systematic search for +1 frameshift sites. Further investigation of the predicted +1 frameshift sequences are in progress. The knowledge of different frameshift sites will enable researchers to better understand translational control.

### **5.10 Acknowledgments**

We acknowledge Robert S. Kuczenski for advice in developing the Python and Matlab program. We are thankful to Dr. Jonathan D. Dinman for his insightful comments of this work.

## REFERENCES

1. Kurland, C.G. (1992) Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.*, **26**, 29-50.
2. Baranov, P.V., Gurvich, O.L., Hammer, A.W., Gesteland, R.F. and Atkins, J.F. (2003) Recode 2003. *Nucleic Acids Res.*, **31**, 87-89.
3. Craigen, W.J. and Caskey, C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, **322**, 273-275.
4. Belcourt, M.F. and Farabaugh, P.J. (1990) Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell*, **62**, 339-352.
5. Farabaugh, P.J., Zhao, H. and Vimaladithan, A. (1993) A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell*, **74**, 93-103.
6. Janetzky, B. and Lehle, L. (1992) Ty4, a new retrotransposon from *Saccharomyces cerevisiae*, flanked by tau-elements. *J. Biol. Chem.*, **267**, 19798-19805.
7. Asakura, T., Sasaki, T., Nagano, F., Satoh, A., Obaishi, H., Nishioka, H., Imamura, H., Hotta, K., Tanaka, K., Nakanishi, H., *et al.* (1998) Isolation and characterization of a novel actin filament-binding protein from *Saccharomyces cerevisiae*. *Oncogene*, **16**, 121-130.
8. Morris, D.K. and Lundblad, V. (1997) Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr. Biol.*, **7**, 969-976.
9. Palanimurugan, R., Scheel, H., Hofmann, K. and Dohmen, R.J. (2004) Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J.*, **23**, 4857-4867.
10. Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F. and Hayashi, S. (1995) Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell*, **80**, 51-60.
11. Shah, A.A., Giddings, M.C., Parvaz, J.B., Gesteland, R.F., Atkins, J.F. and Ivanov, I.P. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**, 1046-1053.
12. Moon, S., Byun, Y., Kim, H.J., Jeong, S. and Han, K. (2004) Predicting genes expressed via -1 and +1 frameshifts. *Nucleic Acids Res.*, **32**, 4884-4892.



13. Ivanov,I.P., Pittman,A.J., Chien,C.B., Gesteland,R.F. and Atkins,J.F. (2007) Novel antizyme gene in *Danio rerio* expressed in brain and retina. *Gene*, **387**, 87-92.
14. Liao,P.Y., Gupta,P., Petrov,A.N., Dinman,J.D. and Lee,K.H. (2008) A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Res.*, **36**, 2619-2629.
15. Anderson,L. and Hunter,C.L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics*, **5**, 573-588.
16. Harrison,P., Kumar,A., Lan,N., Echols,N., Snyder,M. and Gerstein,M. (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.*, **316**, 409-419.
17. Weiss,R.B., Dunn,D.M., Dahlberg,A.E., Atkins,J.F. and Gesteland,R.F. (1988) Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.*, **7**, 1503-1507.
18. Sanders,C.L. and Curran,J.F. (2007) Genetic analysis of the E site during RF2 programmed frameshifting. *RNA*, **13**, 1483-1491.
19. Klump,H.H. (2006) Exploring the energy landscape of the genetic code. *Arch. Biochem. Biophys.*, **453**, 87-92.
20. Curran,J.F. (1993) Analysis of effects of tRNA:Message stability on frameshift frequency at the *Escherichia coli* RF2 programmed frameshift site. *Nucleic Acids Res.*, **21**, 1837-1843.
21. Fluitt,A., Pienaar,E. and Viljoen,H. (2007) Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.*, **31**, 335-346.
22. Jacobs,J.L. and Dinman,J.D. (2004) Systematic analysis of bicistronic reporter assay data. *Nucleic Acids Res.*, **32**, e160.
23. Gupta,P. and Lee,K.H. (2008) Silent mutations result in HlyA hypersecretion by reducing intracellular HlyA protein aggregates. *Biotechnol. Bioeng.*, **101**, 967-974.
24. Kitteringham,N.R., Jenkins,R.E., Lane,C.S., Elliott,V.L. and Park,B.K. (2009) Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. *J. Chromatogr. B. Analyt Technol. Biomed. Life. Sci.*, **877**, 1229-123.
25. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097-6100.

26. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188-1190.
27. Marquez,V., Wilson,D.N., Tate,W.P., Triana-Alonso,F. and Nierhaus,K.H. (2004) Maintaining the ribosomal reading frame: The influence of the E site during translational regulation of release factor 2. *Cell*, **118**, 45-55.
28. Sergiev,P.V., Lesnyak,D.V., Kiparisov,S.V., Burakovsky,D.E., Leonov,A.A., Bogdanov,A.A., Brimacombe,R. and Dontsova,O.A. (2005) Function of the ribosomal E-site: A mutagenesis study. *Nucleic Acids Res.*, **33**, 6048-6056.
29. 29, Nierhaus,K.H. (2006) Decoding errors and the involvement of the E-site. *Biochimie*, **88**, 1013-1019.
30. O'Connor,M., Willis,N.M., Bossi,L., Gesteland,R.F. and Atkins,J.F. (1993) Functional tRNAs with altered 3' ends. *EMBO J.*, **12**, 2559-2566.
31. Lill,R., Lepier,A., Schwagele,F., Sprinzl,M., Vogt,H. and Wintermeyer,W. (1988) Specific recognition of the 3'-terminal adenosine of tRNA<sup>Phe</sup> in the exit site of *Escherichia coli* ribosomes. *J. Mol. Biol.*, **203**, 699-705.
32. Siple,J. and Goldman,E. (1993) Increased ribosomal accuracy increases a programmed translational frameshift in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 2315-2319.
33. Harger,J.W., Meskauskas,A. and Dinman,J.D. (2002) An "integrated model" of programmed ribosomal frameshifting. *Trends Biochem. Sci.*, **27**, 448-454.
34. Pande,S., Vimaladithan,A., Zhao,H. and Farabaugh,P.J. (1995) Pulling the ribosome out of frame by +1 at a programmed frameshift site by cognate binding of aminoacyl-tRNA. *Mol. Cell. Biol.*, **15**, 298-304.
35. Ivanov,I.P., Gurvich,O.L., Gesteland,R.F. and Atkins,J.F. 2003. Recoding: Site- or mRNA-specific alteration of genetic readout utilized for gene expression. In Lapointe,J. and Barker-Gingras,L. (ed.), *Translation mechanism*, Landes Bioscience, Austin, TX, pp. 354-369.
36. Herr,A.J, Wills,N.M., Nelson,C.C., Gesteland,R.F. and Atkins, J.F. (2004) Factors that influence selection of coding resumption sites in translational bypassing: minimal conventional peptidyl-tRNA:mRNA pairing can suffice. *J Biol Chem.*, **279**, 11081-11087.
37. Zaher,H.S. and Green,R. (2009) Quality control by the ribosome following peptide bond formation. *Nature*, **457**, 161-166.

38. Inoue,T., Shingaki,R., Hirose,S., Waki,K., Mori,H. and Fukui,K. (2007) Genome-wide screening of genes required for swarming motility in *Escherichia coli* K-12. *J. Bacteriol.*, **189**, 950-957.
39. Davids,W. and Zhang,Z. (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks--direct evidence of preferential attachment. *BMC Evol. Biol.*, **8**, 23.
40. Fu,C. and Parker,J. (1994) A ribosomal frameshifting error during translation of the *argI* mRNA of *Escherichia coli*. *Mol. Gen. Genet.*, **243**, 434-441.
41. Gurvich,O.L., Baranov,P.V., Zhou,J., Hammer,A.W., Gesteland,R.F. and Atkins,J.F. (2003) Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. *EMBO J.*, **22**, 5941-5950.

## CHAPTER 6

### DIFFERENTIATING TWO TYPES OF THE FRAMESHIFT PROTEINS BY MASS SPECTROMETRY USING MULTIPLE REACTION MONITORING

#### ***6.1 Preface***

This chapter discusses a method to detect the composition of frameshift products for various -1 programmed ribosomal frameshift (PRF) signals. This study will be an accompanying paper to support the experiments in Chapter 7, which will describe a kinetic model for -1 PRF.

#### ***6.2 Abstract***

The -1 programmed ribosomal frameshift (PRF) motif in the human immunodeficiency virus type 1 (HIV-1) genome directs the ribosome to make two types of -1 frameshift proteins: those incorporating the zero frame A-site tRNA and products incorporating the -1 frame A-site tRNA in the recoding site. To date, there has not been a well-established method to quantitatively analyze the two types of frameshift proteins. This study applied nano-flow liquid chromatography electrospray tandem mass spectrometry (nLC-ESI-MS/MS) using multiple reaction monitoring (MRM) to quantify the relative ratio of the two kinds of frameshift proteins. This MS method detected 20.4% and 23.2% of the frameshift proteins incorporating -1 frame A-site tRNAs in the recoding sites of HIV-1 group M type B and group O PRF signals, respectively. A sensitive method to detect the relative population of the frameshift proteins will aid into our understanding of the -1 PRF mechanism.

#### ***6.3 Introduction***

In -1 programmed ribosomal frameshifting (PRF), stimulatory signals in the mRNA

direct the ribosome to translate the -1 reading frame at a certain efficiency. Several viruses, including human immunodeficiency virus type 1 (HIV-1) and the coronavirus for severe acute respiratory syndrome (SARS-CoV), employ -1 PRF to synthesize the enzyme precursors for their replication [1-3]. Importantly, the perturbation of -1 PRF efficiency can damage viral replication (see review by Dinman *et al.*, [4]), which suggests that -1 PRF may serve as a target for antiviral therapeutics.

Viral gene expression involving -1 PRF usually occurs on a heptanucleotide ‘slippery motif’, X XXY YYZ [1,5,6]. A common hypothesis in the field is that P- and A-site tRNAs, decoding XXY and YYZ, reposition with the -1 reading frame to XXX and YYY codons during frameshifting. However, in HIV-1 frameshifting, protein sequencing by Edman degradation revealed that about 70% of the products corresponded to a result of P- and A-site tRNA slippage [1,7]. The other 30% of the products corresponded to PRF that may involve a single P-site tRNA slippage [8]. In Chapter 7, a new kinetic model for -1 PRF is proposed to explain the formation of the two kinds of frameshift products. In SARS-CoV frameshifting, only the product corresponding to a slippage of P- and A-site tRNAs was identified by electrospray mass spectrometry [2]. A recently discovered -1 PRF in alphavirus occurred at a conserved UUUUUUA motif within the coding sequence of 6k [9]. Peptides derived from frameshift products via slippage of P- and A-site tRNAs and a single P-site tRNA slippage were both detected in liquid chromatography tandem mass spectrometry (LC-MS/MS) [9]. However, the ratio between the two products was not determined.

To date, a method to quantitatively analyze the two types of frameshift proteins has not been reported. The two types of -1 PRF proteins differ in only one amino acid in the polypeptide sequences. Therefore, differentiating the two products requires high

mass accuracy. In our previous study, mass spectrometry using MRM successfully detected a peptide derived from a +1 PRF product in *E. coli* [10]. Mass spectrometry using MRM can achieve high specificity and sensitivity by simultaneously monitoring both parent and one or more product ions [11]. The parent ion is the intact analyte and the product ions are generated by the fragmentation of the parent ions. The appropriate parent/product ion pairs, *i.e.* transitions, must be defined or predicted for the analyte of interest. The selected analytes are detected and integrated as peaks in one-dimensional chromatography. Because of the high specificity and sensitivity, MRM assays have been widely applied to the measurement of specific peptides in complex mixtures such as tryptic digests of plasma [11,12].

In this study, FS<sub>0</sub> and FS<sub>-1</sub> denote proteins incorporating the zero frame and the -1 frame A-site tRNA in the recoding motif, respectively. Figure 6.1 summarizes the procedure to detect the fraction of FS<sub>-1</sub> in total frameshift proteins. Purified frameshift proteins containing FS<sub>0</sub> and FS<sub>-1</sub> were digested with trypsin. Peptides spanning the recoding cassette (*i.e.* target peptides) with a single amino acid variation in FS<sub>0</sub> and FS<sub>-1</sub> were quantified by nLC-ESI-MS/MS using MRM. Our result suggested that the MS method is a sensitive approach to detect peptides from frameshift proteins qualitatively and quantitatively.

## **6.4 Materials and methods**

### **6.4.1 Plasmids and bacterial strains**

*Escherichia coli* XL1 blue MRF' (Stratagene, La Jolla, CA) was used in all experimental studies. All constructs were verified by DNA sequencing at the Cornell Bioresource Center. The construction of the dual fluorescence reporter was performed as described previously [13] except that the linker sequences rendered the downstream

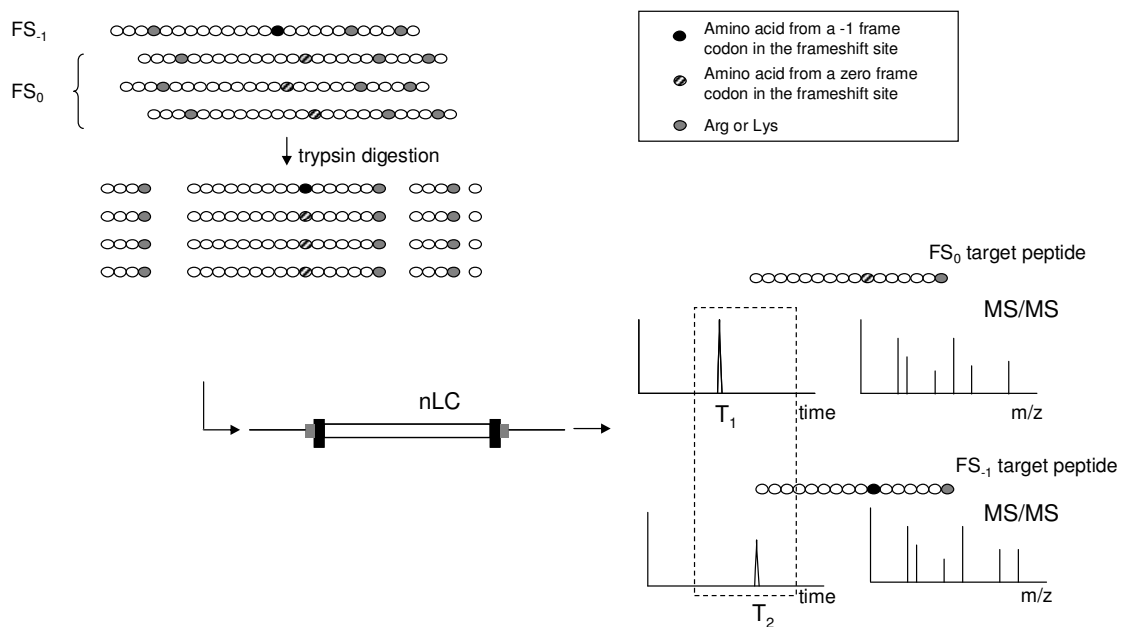


Figure 6.1. Differentiating two types of frameshift products by mass spectrometry.  $FS_0$  and  $FS_{-1}$  represent frameshift proteins incorporating zero frame and -1 frame A-site tRNAs in the frameshift site, respectively. Trypsin cleaves lysine and arginine residues in the polypeptides. Peptides spanning the PRF motif (target peptides) are analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry (nLC-ESI-MS/MS) using multiple reaction monitoring (MRM).

green fluorescence protein (EGFP) coding sequence in the -1 frame. In addition, the reporter plasmid was digested with BsrGI and EcoRI to make the insertion of a 6X histidine tag downstream of *egfp*. The linker sequences were made from complementary oligonucleotides (Integrated DNA Technology, Coralville, IA) and then cloned into SalI and BamHI sites between the coding sequence of DsRed and EGFP in the reporter plasmid. Table 6.1 lists the nucleotide sequences incorporated into the dual fluorescence reporter for testing the compositions of frameshift protein in this study.

#### **6.4.2 Protein sample preparation**

Test strains were grown in 100 ml Luria-Bertani (LB) medium containing 100 µg/ml ampicillin in 500 ml flasks at 250 rpm and 37°C. After 24 hours, 200 OD<sub>600</sub> units of cells were collected by centrifugation at 4000×*g* and 4°C for 20 minutes. Cells were lysed and purified by Ni-NTA under native conditions according to the manufacturer's protocol (Qiagen, Valencia, CA). The purified protein samples were resolved by SDS-PAGE (10% w/v) using Tris-HCl. Gel band excision and in gel trypsin digestion were performed using standard method described previously [14].

#### **6.4.3 Mass spectrometry**

For the standard curves, different concentrations (10 fmol/µl - 0.1 fmol/µl) of a custom synthesized peptide (Bio-Synthesis, Lewisville, TX) solution in 0.1% formic acid (FA) were prepared and 200 nl of each standard solution was injected into Dionex 3000 nLC system (Sunnyvale, CA). First, the sample was loaded in an Acclaim PepMap 100 C18 trap column (300 µm × 5 mm, 5 µm) and on-line desalting was carried out with water (0.1% FA) at a flow rate of 30 µL/min for 5 minutes. Then, peptides trapped in the trap column were delivered to/separated on an Acclaim



Table 6.1 The nucleotide sequences incorporated into the dual fluorescence reporter for testing the compositions of frameshift protein in Chapter 6. The heptanucleotide slippery motifs in the sequence are underlined.

Strain	Plasmid	Linker sequence	PRF cassette origin
MB2	pMB2	TCG ACT GCT AAT <u>TTT TTA</u> GGG AAG ATC TGG CCT TCC TAC AAG GGA AGG CCA GGG AAT TTT CTT GGA TAA AG	HIV-1 group M, type B, <i>gag/pol</i> overlap
O2	pO2	TCG ACT GCT AAT <u>TTT TTA</u> GGG AAG TAC TGG CCT CCG IGG GGC ACG AGG CCA GGC AAT TAT GTG CAG AAA CAA GTG TCC CCA TAA AG	HIV-1 group O <i>gag/pol</i> overlap
P2	pP2	TCG ACT GCC GGT AAG GTG GTC GGT <u>TTT</u> TTG TCG CCG AAC TCG GTG TAA AG	Bacteriophage P2, gene <i>E-E'</i>
PSP3	pPSP3	TCG ACT GCC GCT AAG GTG ATT GGT <u>TTT</u> TTG TCA CCG CTT CGG TAA AG	Bacteriophage PSP3, gene <i>E-E'</i>

PepMap 100 C18 analytical column (75  $\mu\text{m}$   $\times$  15 cm, 3  $\mu\text{m}$ ) with gradients of 2-50% acetonitrile with 0.1% FA over 75 minutes. The eluent was directly introduced into a 4000 QTRAP mass spectrometer through a Nanospray II source (Applied Biosystems, Carlsbad, CA). For MRM, the MIDAS Workflow software (Applied Biosystems) generated a list of possible MRM transitions (Table 6.2) before MS analysis. MS and MS/MS data obtained through MRM were searched using Mascot [15] (v. 2.2, Matrix Science, Boston, MA) within a custom sequence database that included frameshift protein sequences. During the database search, the spectral assignment of MS/MS was performed under parameters of MS tolerance of 1.2 Da, MS/MS tolerance of 0.6 Da, and  $p < 0.05$  and searches were manually confirmed. In addition, peak areas of MRM transitions were calculated using Analyst (v. 1.5, Applied Biosystems). For the test samples, the trypsin digested sample was vacuum dried and reconstituted with 30  $\mu\text{l}$  of 0.1% FA. The analysis was performed as the method for standards except 8  $\mu\text{l}$  of reconstituted samples were used.

## **6.5 Results**

### **6.5.1 Correlation between peptide concentrations and peak area**

Standard peptides with different concentrations were analyzed by nLC-ESI-MS/MS using MRM. These standard peptides mimic the tryptic peptides spanning the PRF motifs in HIV frameshift proteins FS<sub>0</sub> and FS<sub>-1</sub> (Figure 6.2.a). The peak areas correlated well with peptide concentrations for both peptides (Figure 6.2.b and Figure 6.2.c), suggesting that the method is suitable for quantification. For the test samples containing the two peptides in various ratios, the fractions of FS<sub>-1</sub> in total frameshift proteins were calculated in two ways: (1) Peak areas obtained by the MS method were first converted to concentrations by using the standard curves (Figure 6.2). The fraction of FS<sub>-1</sub> was then calculated as the concentration of FS<sub>-1</sub> target peptide over the

Table 6.2 The list of target peptides and their MRM parameters (All peptides except VIGFFVTASVK were confirmed by Mascot search of their MS/MS spectra against a custom-made database (p<0.05)).

Origin	FS <sup>a</sup>	Sequence	Theor MW <sup>b</sup>	Prec $m/z$ <sup>c</sup>	Frag $m/z$ <sup>d</sup>	$z$ <sup>e</sup>	CE <sup>f</sup>
MB2, O2	FS <sub>0</sub>	HSTGAASTANFLR	1331.68	444.90	620.40	3	22.580
MB2, O2	FS <sub>-1</sub>	HSTGAASTANFFR	1365.58	456.20	654.30	3	23.073
P2	FS <sub>0</sub>	VVGFLVAELGVK	1229.73	615.87	199.14	2	35.794
P2	FS <sub>-1</sub>	VVGFFVAELGVK	1263.70	632.86	199.14	2	36.643
PSP3	FS <sub>0</sub>	VIGFLVTASVK	1132.69	567.35	604.36	2	33.370
PSP3	FS <sub>-1</sub>	VIGFFVTASVK	1166.65	584.34	604.36	2	34.217

<sup>a</sup> Type of frameshift protein

<sup>b</sup> Theoretical molecular weight (Da)

<sup>c</sup> Precursor ion  $m/z$

<sup>d</sup> Fragment ion  $m/z$

<sup>e</sup> Charge state

<sup>f</sup> Collision energy (V)

sum of the concentration of FS<sub>0</sub> and FS<sub>-1</sub> target peptides (Eq.1); and (2) the fraction of FS<sub>-1</sub> was directly calculated as the peak area of FS<sub>-1</sub> target peptide divided by the sum of peak areas of FS<sub>-1</sub> and FS<sub>0</sub> target peptides (Eq.2).

$$\text{The fraction of FS}_{-1} \text{ estimated by concentration} = \frac{C_{FS-1}}{C_{FS-1} + C_{FS0}} \times 100\% \quad (\text{Eq.1})$$

where C<sub>FS-1</sub> and C<sub>FS0</sub> are the concentrations, obtained by converting peak areas in MS to concentrations using the standard curve (Figure 6.2), of FS<sub>-1</sub> and FS<sub>0</sub> target peptides in the sample, respectively.

$$\text{The fraction of FS}_{-1} \text{ estimated by peak area} = \frac{A_{FS-1}}{A_{FS-1} + A_{FS0}} \times 100\% \quad (\text{Eq.2})$$

where A<sub>FS-1</sub> and A<sub>FS0</sub> are peak areas for FS<sub>-1</sub> and FS<sub>0</sub> target peptides in MS, respectively.

Figure 6.3 shows a good correlation between the fraction of FS<sub>-1</sub> values calculated by Eq.1 and Eq.2 (R<sup>2</sup>=0.9839). This result suggests that peak area ratios can be used for the relative estimation of FS<sub>-1</sub> fraction in total frameshift proteins. Therefore, Eq.2 was used for later experiments.

### 6.5.2 Detection of two target peptides by nLC-ESI-MS/MS using MRM

Four frameshift signals were investigated for the fraction of FS<sub>-1</sub> in total frameshift proteins (sequences and corresponding strains in Table 6.1): HIV-1 group M type B, HIV-1 group O, bacteriophage P2 and bacteriophage PSP3. In the HIV-1 genome, the PRF cassette located at the *gag/pol* overlap [1]. While the downstream stimulatory signal for the HIV-1 group M type B sequence was suggested to be a stem loop [1], PRF motif in HIV-1 group O may involve a pseudoknot [16]. Bacteriophage P2 is a double stranded DNA phage. Christie *et al.* identified a small reading frame

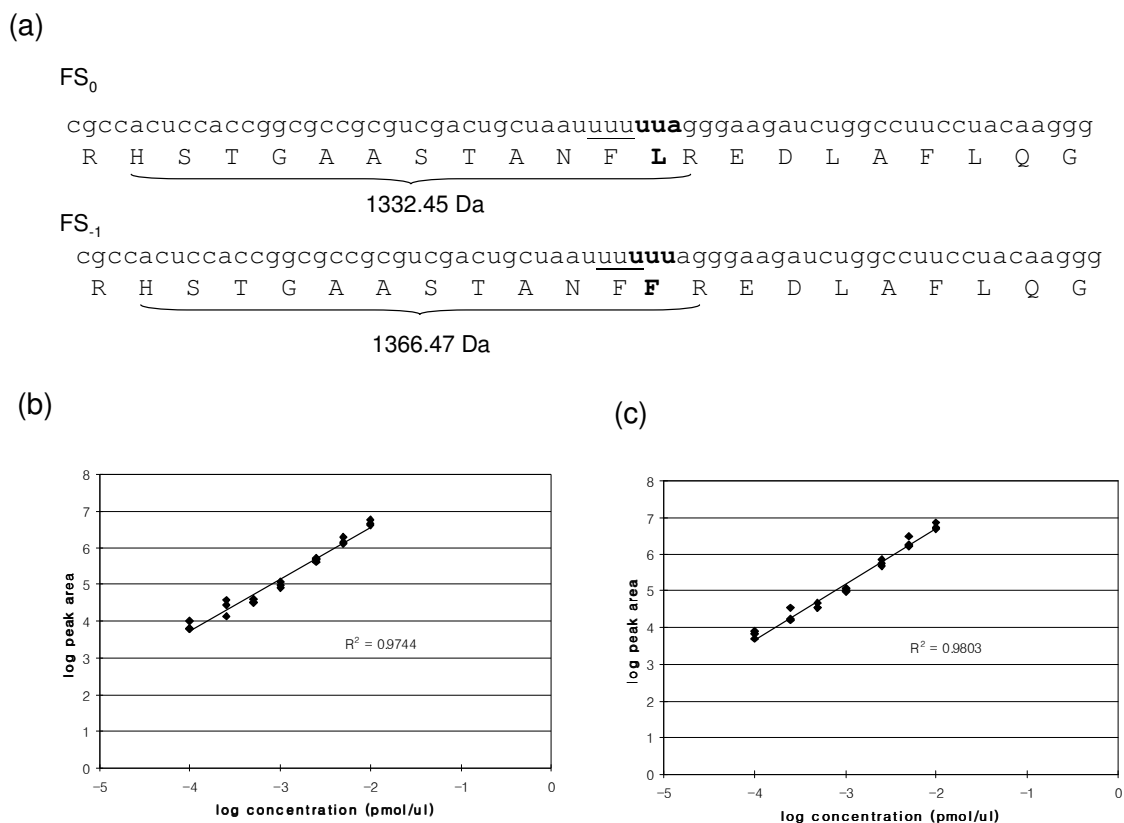


Figure 6.2. Standard curves for HIV-1 target peptides. (a) The mRNA sequence (top) and the polypeptide sequence (bottom) spanning the PRF cassette in HIV-1 group M type B. The P-site in the PRF cassette is underlined. FS<sub>0</sub> incorporates a zero frame A-site tRNA (codon uua for lysine). FS<sub>-1</sub> incorporates a -1 frame A-site tRNA (codon uuu for phenylalanine). Trypsin digestion results in peptides with mass 1332.45 Da and 1366.47 Da from FS<sub>0</sub> and FS<sub>-1</sub>, respectively. (b) Peak areas correlate well with concentrations for HIV-1 FS<sub>0</sub> target peptide HSTGAASTANFLR in log scale. (c) Peak areas correlate well with concentrations for HIV-1 FS<sub>-1</sub> target peptide HSTGAASTANFFR in log scale.

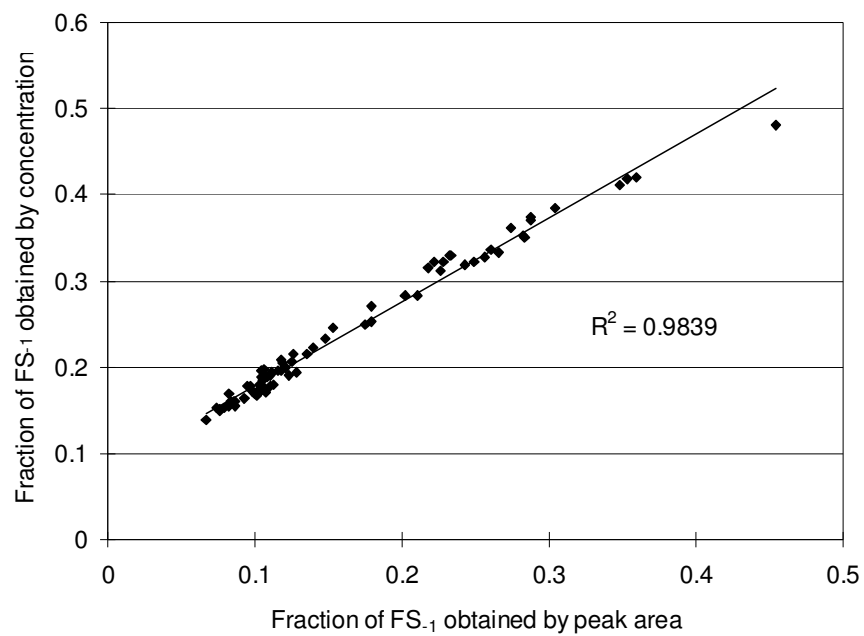


Figure 6.3. Correlation between the fraction of FS-1 values calculated by concentration ratios (Eq.1) and peak area ratios (Eq.2).

overlapping the end of the gene *E* that was translated as a -1 frameshifted extension, namely E+E' [17]. This study observed FS<sub>0</sub> in E+E' frameshifting by Edman degradation, but did not report FS<sub>-1</sub>. In addition, a P2 related phage, PSP3, has a similar genetic structure compared to the P2 frameshift cassette. In the present study, nLC-ESI-MS/MS using MRM detected  $20.4 \pm 1.9\%$  and  $23.2 \pm 1.7\%$  of FS<sub>-1</sub> in total frameshift proteins for MB2 and O2 strains, respectively (Figure 6.4). Interestingly, the method detected 0.85% of FS<sub>-1</sub> in total frameshift proteins for the P2 strain, although this approach did not find FS<sub>-1</sub> for the PSP3 strain.

## 6.6 Discussion

In the present study, nLC-ESI-MS/MS using MRM detected  $20.4 \pm 1.9\%$  and  $23.2 \pm 1.7\%$  of FS<sub>-1</sub> in total frameshift proteins in HIV-1 group M type B and group O derived PRF signals. Our results are consistent with previous studies. Jacks *et al.* performed *in vitro* translation of HIV-1 group M type B frameshift cassette and observed 20-25% of FS<sub>-1</sub> in total frameshift proteins by Edman degradation [1]. Yelverton *et al.* examined the translation of 15-17 nucleotides of HIV-1 frameshift motif in *E. coli* and found about 30% FS<sub>-1</sub> in total frameshift proteins with Edman degradation [7]. The difference in the fraction of FS<sub>-1</sub> in total frameshift proteins may result from different reporter systems being used or different detection methods being employed. In addition, the fraction of FS<sub>-1</sub> in total frameshift proteins observed by Edman degradation was an approximation because: (1) the peak intensity was not consistent for the same amino acid at different positions, suggesting that the peak intensity may not truly reflect the concentration; (2) minor peaks were present in addition to the major peak in one Edman degradation cycle, which may interfere with the detection of the major peak. Because these previous studies did not provide the correlation of the peak height and the concentration of a certain amino acid in the

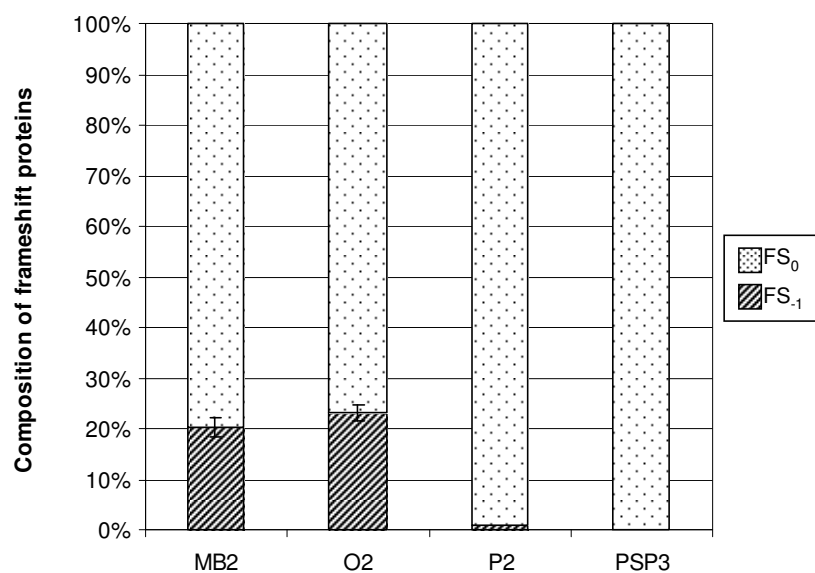


Figure 6.4. Compositions of the frameshift proteins from four PRF cassettes by nLC-ESI-MS/MS using MRM. Error bars denote standard deviations.



peptide, Edman degradation may be less applicable for quantification. On the other hand, nLC-ESI-MS/MS using MRM has been previously used for peptide quantification [18]. Our study observed good correlations for peptide concentrations and peak areas observed in the MS (Figure 6.2). In addition, peptide fractions correlated well with peak area ratios (Figure 6.3). These results suggested that our MS method is suitable for the quantification of the relative population of the frameshift proteins.

The study by Christie *et al.* successfully sequenced frameshift protein E+E' and identified FS<sub>0</sub> in bacteriophage P2 [17]. In our study, 0.85% of FS<sub>-1</sub> in total frameshift proteins was detected in the strain with PRF signal from bacteriophage P2. To our knowledge, this is the first report to demonstrate the presence of FS<sub>-1</sub> in the E+E' frameshifting in bacteriophage P2. FS<sub>-1</sub> was not found in the strain containing PRF motif from the bacteriophage PSP3 genome. It is possible that the PRF motif from the PSP3 genome employs only one mechanism to produce frameshift proteins, or the level of FS<sub>-1</sub> is too low to be detected by our approaches in this case.

To use the peak area ratios obtained from the MS analysis to represent the composition of the frameshift proteins, the following steps were assumed to have the same efficiency for FS<sub>0</sub> and FS<sub>-1</sub>: (1) affinity column purification; (2) enzymatic cleavage of the polypeptide; (3) the release of the peptide from the gel piece (4) ionization of peptides. Additionally, peptides from FS<sub>0</sub> and FS<sub>-1</sub> are assumed to be equally eluted from nLC column. In our reporter system, FS<sub>0</sub> and FS<sub>-1</sub> are about five hundred amino acids in length and they differ only in one amino acid in the linker sequence. In addition, their tryptic peptide spanning the PRF cassettes are the same length. These two factors make these assumptions more likely. If FS<sub>0</sub> and FS<sub>-1</sub> result

in peptides spanning the recoding site with very different length, the assumptions may need to be re-evaluated. Furthermore, the mass spectrometry method requires sequences with a certain level of difference. For example, the frameshift cassette in the *E. coli dnaX* gene is A AAA AAG (where spaces separate the zero frame and the P-site codon is underlined). The zero frame A-site AAG and the -1 frame A-site AAA both code for a lysine, making FS<sub>0</sub> and FS<sub>-1</sub> indistinguishable. Similarly, a leucine/isoleucine variation in FS<sub>0</sub> and FS<sub>-1</sub> are indistinguishable in the MS. Because of the isotopic cluster overlap, for the mass difference only about 1 Da in FS<sub>0</sub> and FS<sub>-1</sub>, *e.g.* glutamic acid (M.W. 129.12) and lysine (M.W. 128.17), our approach is also less applicable.

## 6.7 Conclusion

This study applied nLC-ESI-MS/MS using MRM for the relative quantification of the ratio of the two kinds of frameshift proteins derived from four PRF cassettes: HIV-1 group M type B, HIV-1 group O, bacteriophage P2 and bacteriophage PSP3. This method detected that  $20.4 \pm 1.9\%$  and  $23.2 \pm 1.7\%$  of frameshift proteins incorporated -1 frame A-site tRNAs in HIV-1 group M type B and group O derived PRF signals, respectively. For the bacteriophage P2 frameshift cassette, 0.85% of frameshift proteins incorporated -1 frame A-site tRNAs. This approach detected no frameshift proteins incorporating the -1 frame A-site tRNAs in bacteriophage PSP3 frameshift site. This MS method provides a platform to investigate frameshift protein production through different mechanisms.

## REFERENCES

1. Jacks,T., Power,M.D., Masiarz,F.R., Luciw,P.A., Barr,P.J. and Varmus,H.E. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280-283.
2. Baranov,P.V., Henderson,C.M., Anderson,C.B., Gesteland,R.F., Atkins,J.F. and Howard,M.T. (2005) Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology*, **332**, 498-510.
3. Brierley,I. and Dos Ramos,F.J. (2006) Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. *Virus Res.*, **119**, 29-42.
4. Dinman,J.D., Ruiz-Echevarria,M.J. and Peltz,S.W. (1998) Translating old drugs into new treatments: Ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.*, **16**, 190-196.
5. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2002) Recoding: Translational bifurcations in gene expression. *Gene*, **286**, 187-201.
6. Namy,O., Rousset,J.P., Naphine,S. and Brierley,I. (2004) Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell*, **13**, 157-168.
7. Yelverton,E., Lindsley,D., Yamauchi,P. and Gallant,J.A. (1994) The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in *Escherichia coli*. *Mol. Microbiol.*, **11**, 303-313.
8. Horsfield,J.A., Wilson,D.N., Mannering,S.A., Adamski,F.M. and Tate,W.P. (1995) Prokaryotic ribosomes recode the HIV-1 gag-pol-1 frameshift sequence by an E/P site post-translocation simultaneous slippage mechanism. *Nucleic Acids Res.*, **23**, 1487-1494.
9. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2004) P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221-230.
10. Firth,A.E., Chung,B.Y., Fleeton,M.N. and Atkins,J.F. (2008) Discovery of frameshifting in alphavirus 6K resolves a 20-year enigma. *Virol. J.*, **5**, 108.
11. Liao,P.Y., Choi,Y.S. and Lee,K.H. (2009) FSscan: A mechanism-based program to identify +1 ribosomal frameshift hotspots. *Nucleic Acids Res.*, doi:10.1093/nar/gkp796.
12. Anderson,L. and Hunter,C.L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics*, **5**, 573-588.

13. Kuhn,E., Wu,J., Karl,J., Liao,H., Zolg,W. and Guild,B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and <sup>13</sup>C-labeled peptide standards. *Proteomics*, **4**, 1175-1186.
14. Liao,P.Y., Gupta,P., Petrov,A.N., Dinman,J.D. and Lee,K.H. (2008) A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Res.*, **36**, 2619-2629.
15. Finehout,E.J. and Lee,K.H. (2003) Comparison of automated in-gel digest methods for femtomole level samples. *Electrophoresis*, **24**, 3508-3516.
16. Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.
17. Baril,M., Dulude,D., Steinberg,S.V. and Brakier-Gingras,L. (2003) The frameshift stimulatory signal of human immunodeficiency virus type 1 group O is a pseudoknot. *J. Mol. Biol.*, **331**, 571-583.
18. Christie,G.E., Temple,L.M., Bartlett,B.A. and Goodwin,T.S. (2002) Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. *J. Bacteriol.*, **184**, 6522-6531.
19. Lange,V., Picotti,P., Domon,B. and Aebersold,R. (2008) Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol. Syst. Biol.*, **4**, 222. doi:10.1038/msb.2008.61.

## CHAPTER 7

### KINETIC MODEL ANALYSIS OF THE EFFECT OF DIFFERENT ELONGATION STEPS ON -1 RIBOSOMAL FRAMESHIFTING

#### **7.1 Preface**

Chapter 4 describes a kinetic model for +1 programmed ribosomal frameshifting (PRF) which reveals the effects of different interactions on +1 PRF. A similar approach is applied to understating the mechanism of -1 PRF. This chapter presents a kinetic model for -1 PRF. The model predictions are tested experimentally using a dual fluorescence reporter system in *Escherichia coli*. The results suggest that several steps in the translation elongation have more prominent effects on -1 PRF efficiency and the percentage of the two types of -1 frameshift products.

#### **7.2 Abstract**

Several important viruses including the human immunodeficiency virus type 1 (HIV-1) and the coronavirus for severe acute respiratory syndrome (SARS-CoV) employ -1 programmed ribosomal frameshifting (PRF) for their protein expression. Here, a kinetic framework is developed to describe -1 PRF. The model yields two possible -1 frameshift products: those incorporating zero frame A-site tRNAs in the recoding site and products incorporating -1 frame A-site tRNAs in the recoding site. Using known kinetic rate constants, the individual contributions of different translation steps to -1 PRF and the ratio between two types of frameshift products are evaluated. A dual fluorescence reporter system is employed in *Escherichia coli* to empirically test the model. In addition, the study presents a novel mass spectrometry approach to quantify the ratio of the two frameshift products. A more detailed understanding of -1 PRF mechanism may provide insight into developing antiviral therapeutics.

### 7.3 Introduction

Programmed ribosomal frameshifting (PRF) is a process where specific signals in the mRNA direct the ribosome to switch the reading frame at a certain efficiency. In -1 PRF, the ribosome slips one nucleotide towards the 5'-end of the mRNA during translation. Several viruses, including human immunodeficiency virus type 1 (HIV-1) and the coronavirus for severe acute respiratory syndrome (SARS-CoV), employ -1 PRF to synthesize precursors of enzymes for their replication [1,2] and the ratio of the zero frame and -1 frame products is important to the vitality of the organism [3,4]. As such, altering -1 PRF efficiency may damage viral replication (see review by Dinman *et al.*, [5]), which suggests that -1 PRF can serve as a target for the development of antiviral therapeutic.

-1 programmed ribosomal frameshifting usually consists of three essential mRNA elements: (1) a 'slippery' heptanucleotide X XXY YYZ (X can be any nucleotide, Y is A or U and Z is not G in eukaryotes; spaces separate the initial reading frame), where the ribosome changes the reading frame [3,6]; (2) a downstream stimulatory mRNA secondary structure [7-9]; and (3) a spacer between the slippery sequence and the stimulatory signal. It has been suggested that the stimulatory signal promotes -1 PRF efficiency by making the ribosome pause over the slippery sequence [10-12]. The length of the spacer has also been shown to affect frameshift efficiency [6,8,13].

As PRF occurs during translation elongation, the models of -1 PRF should be described within this context. The elongation cycle can be divided into four stages. First, the ribosome selects the cognate aminoacyl-tRNA (aa-tRNA) according to the codon at the decoding center (decoding, DC in Figure 7.1). Second, the aa-tRNA moves from A/T entry state into the A/A state to be accommodated into the ribosome

(aa-tRNA accommodation, AA in Figure 7.1). Third, the ribosome catalyzes the peptidyl transfer, resulting in a peptidyl tRNA in the A-site and a deacylated tRNA in the P site (peptidyl transfer, PT in Figure 7.1). Fourth, the peptidyl-tRNA moves from the A-site to the P-site, carrying the mRNA along, and the deacylated tRNA moves out of the P-site into the E-site from where it dissociates (translocation, TL in Figure 7.1). Translocation opens up the ribosome A-site and the ribosome moves on to another aa-tRNA selection.

Two major hypotheses have been proposed for the mechanism of -1 PRF. One hypothesis proposes that -1 PRF takes place during the accommodation of the aa-tRNA [7,14,15]. The simultaneous-slippage model by Jacks *et al.* [7] suggests that peptidyl- and aa-tRNAs slipped to base pair with the -1 reading frame during aa-tRNA accommodation. In the model by Farabaugh [14], -1 PRF occurs when aa-tRNA and peptidyl-tRNA are located in the A/T entry and P/P site. In agreement with these models, Plant *et al.*, [15] further proposed a 9Å model for -1 PRF. During the accommodation, the movement of the anticodon loop of the aa-tRNA is about 9 Å [16]. The 9Å model suggests that the movement of 9Å by the anticodon loop is constrained in the presence of the stimulatory RNA, causing a tension that can be relieved by -1 slippage of the tRNA. Consistently, mutations altering aa-tRNA accommodation were found to affect -1 PRF [17-19]. However, these models do not explain the role of the sequence upstream of the slippery site, which is shown to affect the -1 PRF efficiency [20,21]. The second hypothesis proposes that -1 PRF occurs during translocation. Weiss *et al.* [22] suggests that after peptidyl transfer, the two tRNAs move to P/E and A/P states, where the two tRNAs may dissociate from the mRNA and re-pair with the new reading frame. Cryoelectron microscopy imaging reveals that a pseudoknot interacts with the ribosome to block the mRNA entrance





channel, compromising the translocation process during -1 PRF [23]. In the model by Leger *et al.* [21], -1 PRF is triggered by an incomplete translocation and ribosome E-, P- and A-sites are all involved in the process. Their model suggests that in the presence of a stimulatory signal, the translocation is incomplete and a transition intermediate is formed. The entry of the new aa-tRNA into the ribosome and the tendency of tRNAs to revert to stable states drives the shift of the reading frame. The model is supported by evidence that mutations altering E-site tRNA binding affect -1 PRF [21]. However, these incomplete translocation models do not explain two types of frameshift proteins being found in HIV-1 frameshifting, which will be described next.

Protein sequencing has confirmed the -1 PRF site for HIV-1 is U UUU UUA (where the P-site of the ribosome during frameshifting is underlined), located within the gag/pol overlap [1]. Interestingly, about 70% of the frameshift products contain Phe-Leu (derived from UUU UUA decoding) and 30% of the products contain Phe-Phe (derived from UUU UUU decoding) at the frameshift site [1,24]. Previous studies suggested that the product with Phe-Phe at the frameshift site may result from a single P-site tRNA slippage [1,24,25]. Consequently, the -1 frame aa-tRNA is recruited to the ribosome. However, the precise mechanism that drives the process is not clear. To date, no model has been proposed to explain the formation of different frameshift proteins simultaneously. Here, we develop a kinetic model for -1 PRF to explain previous experimental observations, reveal major steps in translation elongation that affect -1 PRF, and reconcile various models for -1 PRF discussed in the literature. In addition, -1 PRF efficiency was tested *in vivo* by a dual fluorescence reporter and the composition of different frameshift proteins was analyzed by mass spectrometry. In agreement with the model predictions, the experimental perturbation of different steps

in translation results in different levels of -1 PRF efficiency as well as the composition of two types of frameshift proteins.

#### ***7.4 Kinetic model***

In our previous study, a kinetic model successfully presented the effect of ribosome E-, P-, and A-site interactions on +1 PRF [26]. A similar approach can be applied to understanding the mechanism of -1 PRF. The mechanistic model in the present study proposes that -1 PRF can occur during translocation and aa-tRNA accommodation. When -1 PRF occurs during translocation, the ribosome moves two, rather than three, bases toward the 3' end of the mRNA, thus shifting the reading frame. In this case, a -1 frame codon will be present in the A-site. Consequently, the -1 frame translation starts at the A-site of the recoding site. When -1 PRF occurs during aa-tRNA accommodation, the two tRNAs in the ribosome P- and A- sites slip to base pair with the -1 reading frame. Consequently, the -1 frame aa-tRNA incorporation starts one codon downstream of the frameshift site (Figure 7.1).

An elegant series of biochemical analyses have established detailed kinetic models of translocation [27] and aa-tRNA selection [28]. Translocation involves EF-G binding to the pretranslocational ribosome, GTP hydrolysis, unlocking conformation change, Pi release, tRNA movement, relocking conformation change, and dissociation of EF-G from the posttranslocational ribosome. This concept is illustrated along the top of Figure 7.2 from component PA (pretranslocational ribosome) to  $E_0P_0$  (posttranslocational ribosome). The selection and accommodation of aa-tRNA involves initial binding of the ternary complex EF-Tu:aa-tRNA:GTP, codon recognition, EF-Tu GTPase activation, GTP hydrolysis, dissociation of EF-Tu from the ribosome, and accommodation of the acceptor end of the aa-tRNA into the A-site

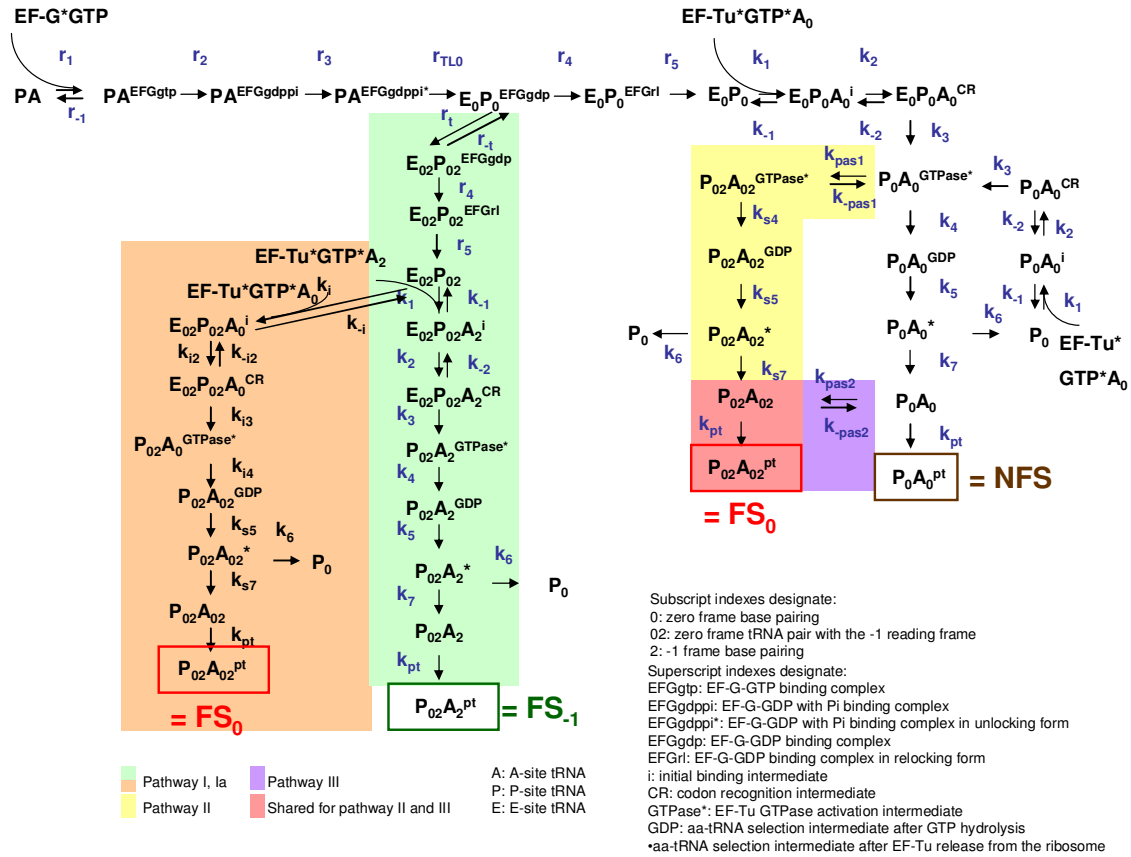


Figure 7.2. The kinetic framework for -1 PRF. Top: the procedure from PA to  $E_0P_0$  represents translocation, which involves reversible EF-G binding ( $r_1$ ,  $r_{-1}$ ), GTP hydrolysis ( $r_2$ ), unlocking conformation change ( $r_3$ ), tRNA movement and Pi release ( $r_{TL0}$ ), re-locking conformation change ( $r_4$ ), and EFGP dissociation ( $r_5$ ). The  $E_0P_0$  complex then undergoes aa-tRNA selection: from  $E_0P_0$  to  $P_0A_0$ . The selection of aa-tRNA involves: reversible EF-Tu binding ( $k_1$ ,  $k_{-1}$ ), reversible codon recognition ( $k_2$ ,  $k_{-2}$ ), GTPase activation ( $k_3$ ), GTP hydrolysis ( $k_4$ ), EF-Tu conformation change and dissociation ( $k_5$ ), aa-tRNA rejection by proofreading ( $k_6$ ) or aa-tRNA accommodation ( $k_7$ ). The elongation cycle without any PRF event will result in forming non-frameshift proteins (NFS). Pathway I in green suggests that -1 PRF occurs during the reloading step in translocation, which leads to the formation of  $FS_{-1}$ . Pathway Ia indicates the  $E_{02}P_{02}$  complex may interact with a zero frame aa-tRNA and eventually produce  $FS_0$ . Pathways II and III suggests that -1 PRF occurs during aa-tRNA selection and accommodation, resulting in producing  $FS_0$ .

or the rejection of the aa- tRNA by proofreading. These aa-tRNA selection steps are illustrated in the process from  $E_0P_0$  to NFS in Figure 7.2.

Our kinetic model suggests three reaction pathways that generate -1 frameshift proteins (Figure 7.2). In Pathway I, a secondary structure blocks the entrance of the mRNA channel and induces -1 PRF during translocation. Specifically, the shift of the reading frame occurs between the tRNA movement and Pi release (rate constant  $r_{TL0}$ ) and the relocking step (rate constant  $r_4$ ). Weiss *et al.* [22] suggested that when the two tRNAs move from P/E and A/P to E/E and P/P states, they can un-pair from the mRNA and re-pair with the -1 reading frame. In our model,  $r_t$  represents the rate constant for a ribosome:EF-G:GDP complex with two tRNAs in the E- and P-sites ( $E_0P_0^{EFGgdp}$ ) to re-pair with the -1 reading frame ( $E_{02}P_{02}^{EFGgdp}$ ). This motion is reversible, and  $r_{-t}$  denotes the rate constant for the reverse reaction. This step is followed by a relocking conformation change and EF-G release from the ribosome complex. The resulting  $E_0P_0$  and  $E_{02}P_{02}$  will then move on to the aa-tRNA selection, according the codon presented in the empty A-site. Here,  $E_0P_0$  and  $E_{02}P_{02}$  are ribosomes with E- and P-sites occupied without EF-G binding, where subscript 0 means a zero frame tRNA pairs with the zero frame; subscript 02 means a zero frame tRNA pairs with the -1 frame.  $E_0P_0$  may generate non-frameshift product NFS, or enter Pathway II or III described below.  $E_{02}P_{02}$  can generate frameshift product FS<sub>-1</sub>, which incorporates a -1 frame aa-tRNA in the frameshift site. In addition, it is also possible that the ribosome complex  $E_{02}P_{02}$  recruits a zero frame aa-tRNA ( $A_0$ ) and accommodates this aa-tRNA into the -1 frame. In this case, frameshift product FS<sub>0</sub>, which incorporates a zero frame aa-tRNA in the frameshift site, is produced (Pathway Ia).

In the second and third pathways, a secondary structure slows down the ribosome movement and induces -1 PRF during aa-tRNA accommodation. Pathway II suggests that a slippage of P- and A-site tRNAs occurs before GTP hydrolysis. In Figure 7.2, the process from  $P_0A_0^{GTPase*}$  to  $P_{02}A_{02}^{GTPase*}$  with the rate constant  $k_{pas1}$  describes the slippage in Pathway II. The ribosome complex ( $P_{02}A_{02}^{GTPase*}$ ) will proceed to the remaining steps of the aa-tRNA selection to reach the peptidyl transfer step.

Therefore, Pathway II can generate  $FS_0$ . Pathway III suggests that a slippage of P- and A-site tRNAs occurs before peptidyl transfer. In Figure 7.2, the process from  $P_0A_0$  to  $P_{02}A_{02}$  with the rate constant  $k_{pas2}$  describes the slippage in Pathway III.  $P_0A_0$  and  $P_{02}A_{02}$  then go through peptidyl transfer, generating NSF and  $FS_0$ , respectively.

## **7.5 Materials and methods**

### **7.5.1 Computation of the kinetic model**

All pathways were mathematically described as systems of ordinary differential equations (see Supplementary data). Assuming steady state, the expressions of intermediate concentrations in terms of initial reactant (PA) were solved by Matlab v.R2008a (Mathworks Inc., Natick, MA). By applying the empirically-determined rate constants and assumed ranges of rate constants of incomplete translocation, P- and A-site tRNA slippage (Table 7.S1-7.S4 in Supplementary data), the amount of non-frameshift proteins NFS ( $P_0A_0^{pt}$  in the kinetic model) and two types of frameshift proteins,  $FS_{-1}$  ( $P_{02}A_2^{pt}$  in the kinetic model) and  $FS_0$  ( $P_{02}A_{02}^{pt}$  in the kinetic model) were calculated. The frameshift efficiency (FS%) in the model is defined as the amount of frameshift proteins divided by the amount of total proteins and multiplied by 100 % (Eq.1). The fraction of  $FS_{-1}$  is calculated as the amount of  $FS_{-1}$  divided by the amount of total frameshift proteins and multiplied by 100 % (Eq.2).

$$FS\% = \frac{(FS_{-1} + FS_0)}{(NFS + FS_{-1} + FS_0)} \times 100\% \quad \text{Eq.1}$$

$$\text{Fraction of } FS_{-1}(\%) = \frac{(FS_{-1})}{(FS_{-1} + FS_0)} \times 100\% \quad \text{Eq.2}$$

### 7.5.2 Plasmids and bacterial strains

*Escherichia coli* XL1 blue MRF' (Stratagene, La Jolla, CA) was used in all experimental studies. All constructs were verified by DNA sequencing at the Cornell Bioresource Center. The construction of the dual fluorescence reporter was performed as described previously [26,29] and in Chapter 6, except that different linker sequences were incorporated into the reporter plasmid (Table 7.1). These sequences are derived from the frameshift signal in HIV-1 group M subtype B [21]. For MB2 strain, the linker sequence was made from complementary oligonucleotides (Integrated DNA Technology, Coralville, IA) and then cloned into Sall and BamHI sites between the coding sequence of DsRed and EGFP in the reporter plasmid. For the MB2CCC strain, the nucleotide sequence was mutated by site-directed mutagenesis according to the manufacturer's protocol (Qiagen, Valencia, CA).

### 7.5.3 *In vivo* fluorescence assay

Cells with the appropriate plasmids were cultured in 1 ml Luria-Bertani (LB) medium containing 100 µg/ml ampicillin and with or without 0.75 µg/ml chloramphenicol in a 24-well plate for 24 hours at 250 rpm and 37°C. The fluorescence was then measured by a plate reader (SpectraMax M5, Molecular Devices, Sunnyvale, CA). The fluorescence measurement was performed as described previously [26,29]. Experimental frameshift efficiency (FS%<sub>exp</sub>) was obtained as the ratio of green fluorescence to red fluorescence for the test strains, normalized against the

Table 7.1 Linker sequences and corresponding *E. coli* strains in Chapter 7. The heptanucleotide slippery motifs in the sequence are underlined.

Linker sequence between the two fluorescence reporter coding sequence	Strain
GCT AAT <u>TTT TTA</u> GGG AAG ATC TGG CCT TCC TAC AAG GGA AGG CCA GGG AAT TTT CTT GGA TAA AG	MB2
GCC CCT <u>TTT TTA</u> GGG AAG ATC TGG CCT TCC TAC AAG GGA AGG CCA GGG AAT TTT CTT GGA TAA AG	MB2CCC

fluorescence ratio of the control strain. Statistical analysis was applied to all datasets according to Jacobs *et al.* [30]. Twenty-three to forty-six replicates for test strains and control strains were performed to satisfy the minimum sample requirement for statistical significance.

#### **7.5.4 Protein purification and trypsin digestion**

The method for protein purification and trypsin digestion was performed as described in Chapter 6.

#### **7.5.5 Mass spectrometry analysis**

Protein samples were analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry (nLC-ESI-MS/MS) using multiple reaction monitoring (MRM) as described in Chapter 6. Table 7.S5 lists the parameters for MRM.

### **7.6 Results**

#### **7.6.1 Mathematical model**

The kinetic model allows for the evaluation of the effect of different translation steps on FS% and the fraction of FS<sub>-1</sub>. In addition, sensitivity analysis reveals several parameters that have a greater influence on FS% (Supplementary data). Therefore, the model results will focus on these higher impact parameters.

In Pathway I, -1 PRF occurs during translocation. Two parameters play important roles in Pathway I in the kinetic model. In the model,  $r_t$  represents the rate constant for incomplete translocation. An increase in  $r_t$  while other parameters in the model remain constants leads to an increase in FS% (blue line in Figure 7.3.a). Both the levels of FS<sub>-1</sub> and FS<sub>0</sub> increase when  $r_t$  increases (green and red lines in Figure 7.3.a). Because



the rise in the  $FS_{-1}$  level is larger, increasing  $r_t$  results in a larger  $FS_{-1}$  fraction (Figure 7.3.b). The rate constant  $r_4$  accounts for the relocking step during translocation. A decrease in  $r_4$  while other parameters in the model remain constant results in an increase in  $FS\%$  (blue line in Figure 7.4.a). Both the levels of  $FS_{-1}$  and  $FS_0$  increase when  $r_4$  decreases (green and red lines in Figure 7.4.a). The increase in the  $FS_{-1}$  level is larger, leading to a larger  $FS_{-1}$  fraction with a decrease in  $r_4$ . The results suggest that translocation perturbations by a downstream mRNA secondary structure, mutations or chemical inhibitors may result in a higher  $FS\%$ , primarily due to a larger amount of  $FS_{-1}$  being produced. Notably, manipulating  $r_t$  values causes larger changes in  $FS\%$  and the  $FS_{-1}$  fraction compared to the effect of  $r_4$ , suggesting a dominant role of  $r_t$  on -1 PRF in Pathway I. Previous experimental studies observed that mutating the E-site codon in the recoding site and the presence of a translocation inhibitor altered  $FS\%$  [21] which is consistent with our observations.

In Pathways II and III, -1 PRF occurs during aa-tRNA accommodation. In Pathway II, the slippage of P- and A-site tRNAs occurs before GTD hydrolysis, while in Pathway III, the slippage occurs before peptidyl transfer. Sensitivity analysis reveals that the parameters in Pathway II have relatively little impact on  $FS\%$ . On the other hand, the analysis shows that the rate constant  $k_{pas2}$ , representing the slippage in Pathway III, has a more significant impact on  $FS\%$ . Figure 7.5.a shows that a higher  $k_{pas2}$  results in a higher  $FS\%$ . Interestingly, the larger  $FS\%$  results from an increase in  $FS_0$  while the level of  $FS_{-1}$  remains at a similar level (Figure 7.5.a). Therefore, the fraction of  $FS_{-1}$  decreases when  $k_{pas2}$  increases (Figure 7.5.b).

In the model,  $k_{pt}$  represents the rate constant for peptidyl transfer, which is the last step of all three pathways. The model predicts that a decrease in  $k_{pt}$  would result in a higher

FS% and FS<sub>0</sub>, and a similar level of FS<sub>-1</sub> (Figure 7.6.a). Consequently, a smaller fraction of FS<sub>-1</sub> is observed when  $k_{pt}$  decreases (Figure 7.6.b). The model results are consistent with previous experimental observations that peptidyl transferase inhibitors affect FS% [31].

### 7.6.2 Experimental results

To examine the model predictions, the frameshift efficiency was tested *in vivo* using a dual fluorescence reporter system. In addition, the composition of frameshift proteins was analyzed by mass spectrometry. A more detailed evaluation of the MS method for frameshift protein analysis is described in Chapter 6.

Chloramphenicol can inhibit peptidyl transfer during translation [32]. Addition of this drug into the culture should decrease the rate of peptidyl transfer, *i.e.* a decrease in  $k_{pt}$  in the model. The model predicts that a smaller  $k_{pt}$  causes a higher FS% and a lower fraction of FS<sub>-1</sub> (Figure 7.6). Consistently, a 2.1-fold increase in FS%<sub>exp</sub> is observed in the *E. coli* culture with 0.75 µg/ml chloramphenicol compared to the culture without the drug. The fractions of FS<sub>-1</sub> for the culture with and without chloramphenicol are 17.3% and 20.4%, respectively (Figure 7.7.a). Although a slight decrease in the fraction of FS<sub>-1</sub> is observed in the presence of the drug, the difference is not statistically significant ( $p>0.05$ ).

The frameshift sequence for HIV-1 is U AAU UUU UUA, where a space separates each zero frame codon and the P-site is underlined. The E-site tRNA<sub>G<sup>U</sup>U</sub><sup>Asn</sup> may form one canonical base pairing with the -1 frame UAA. In the MB2CCC strain, the sequence was mutated to **C** **CCU** UUU UUA (mutations shown in bold). The E-site tRNA<sub>G<sup>G</sup>G</sub><sup>Pro</sup> can form three canonical base pairings with the -1 frame CCC. Because

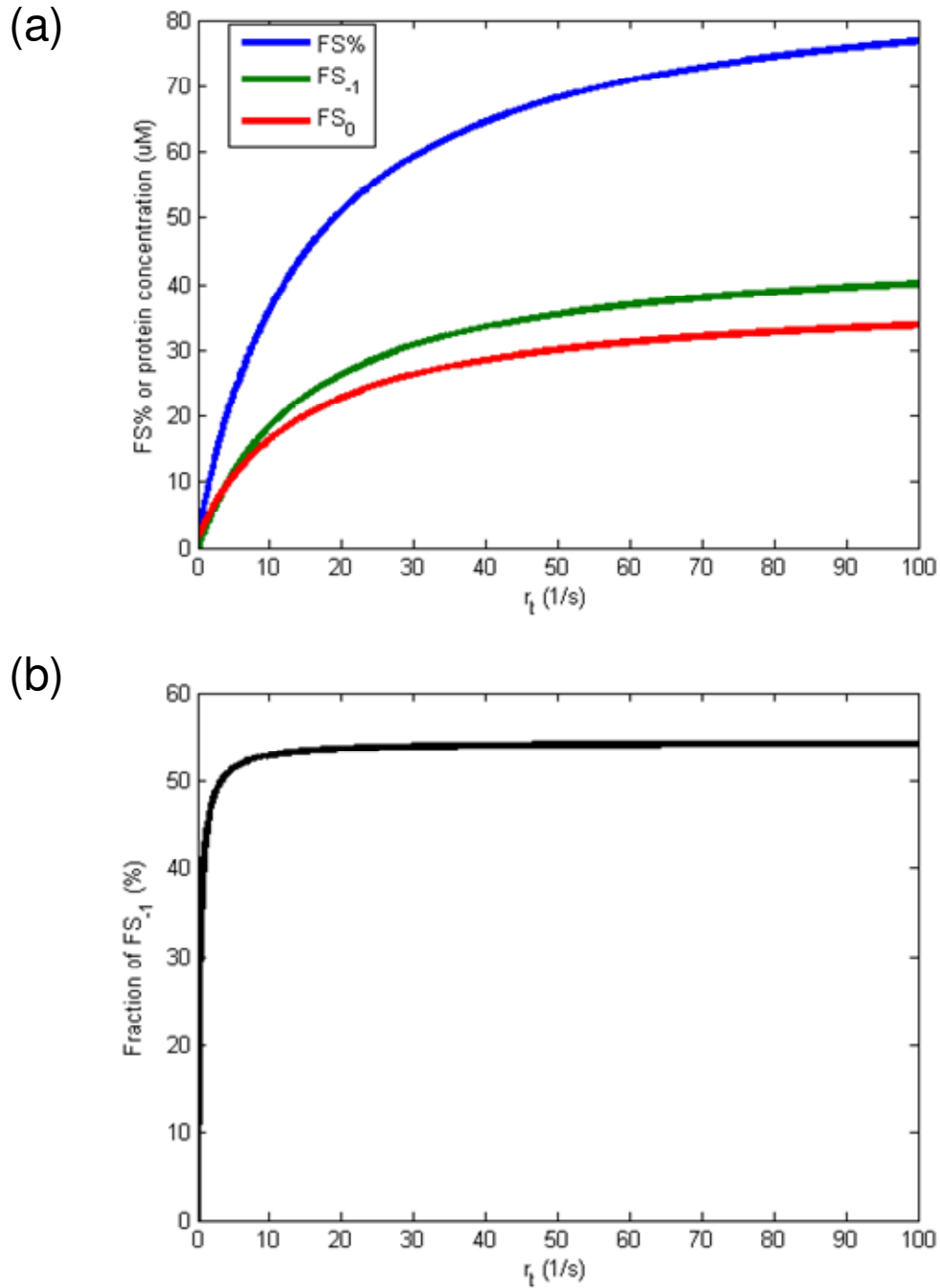


Figure 7.3. The effect of incomplete translocation (represented by  $r_t$ ) on -1 PRF. All the other parameters are assumed to be constant. (a) The effect of  $r_t$  on the level of frameshift efficiency (FS%, blue line), frameshift protein incorporating a -1 frame aa-tRNA at the recoding site (FS<sub>-1</sub>, green line) and frameshift protein incorporating a zero frame aa-tRNA at the recoding site (FS<sub>0</sub>, red line). (b) The effect of  $r_t$  on the fraction of FS<sub>-1</sub>.

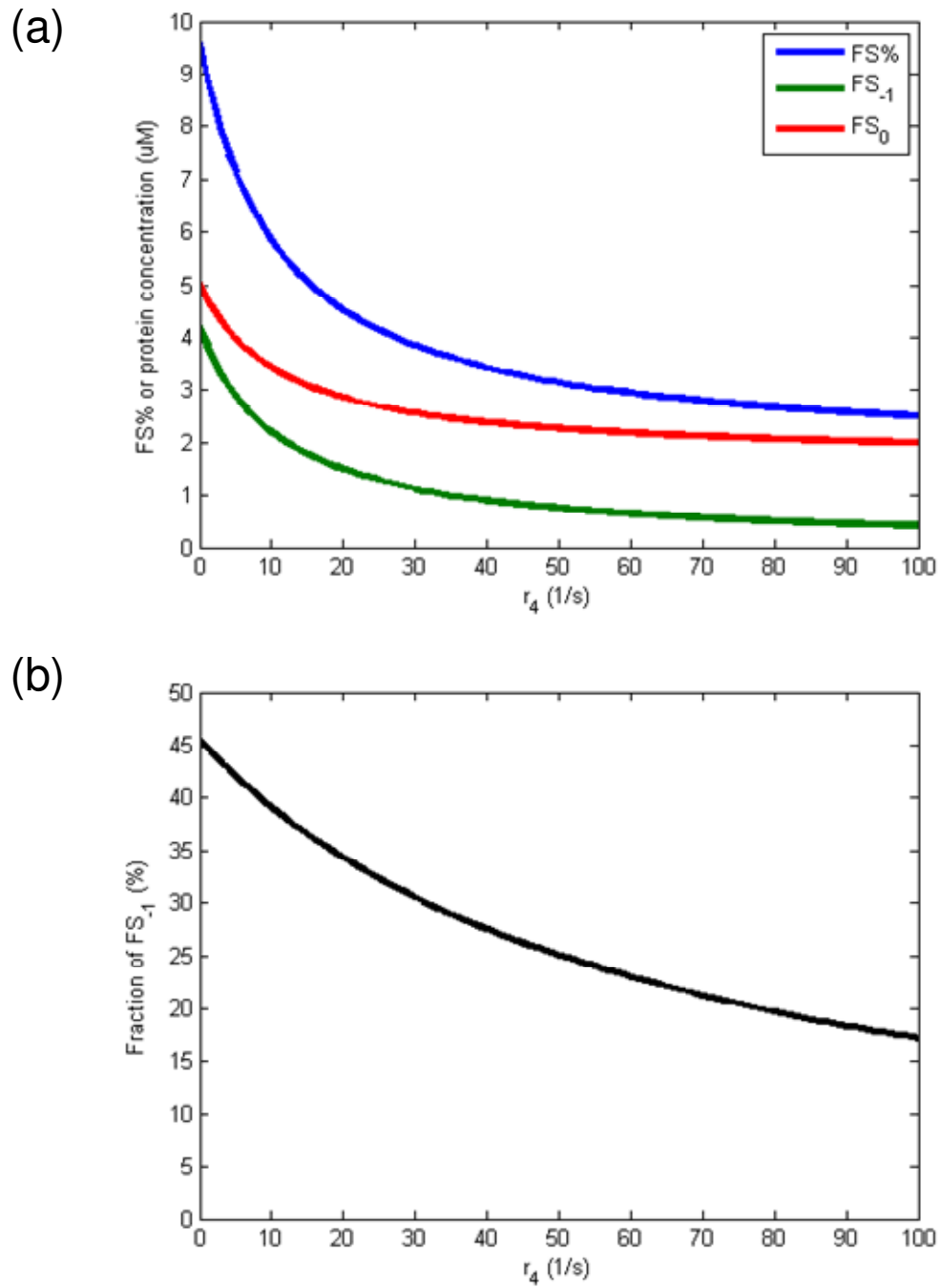


Figure 7.4. The effect of the relocking step during translocation (represented by  $r_4$ ) on -1 PRF. All the other parameters are assumed to be constant. (a) The effect of  $r_4$  on the level of frameshift efficiency (FS%, blue line), frameshift protein incorporating a -1 frame aa-tRNA at the recoding site (FS<sub>-1</sub>, green line) and frameshift protein incorporating a zero frame aa-tRNA at the recoding site (FS<sub>0</sub>, red line). (b) The effect of  $r_4$  on the fraction of FS<sub>-1</sub>.

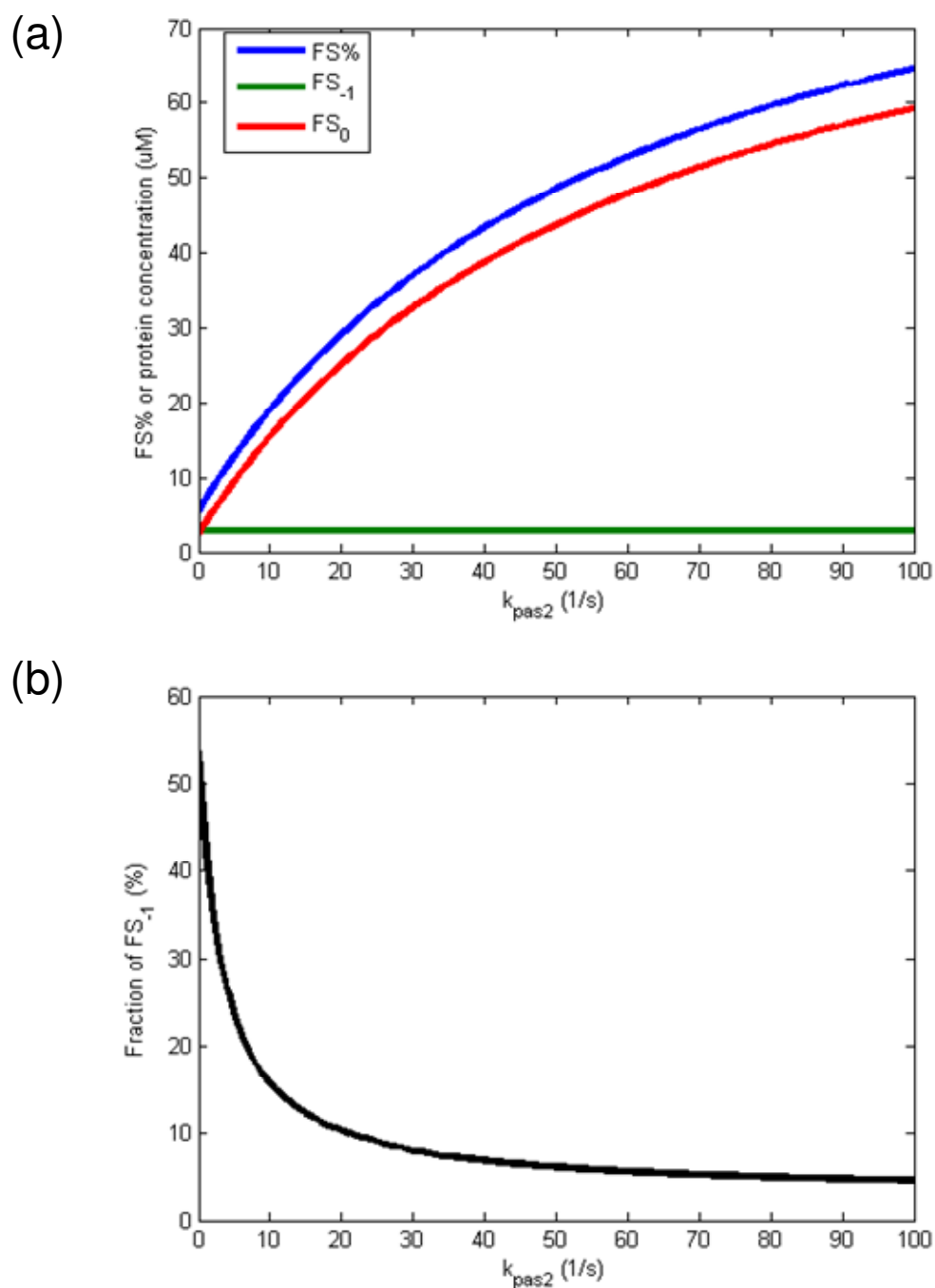


Figure 7.5. The effect of the slippage of P- and A-site tRNAs before peptidyl transfer (represented by  $k_{pas2}$ ) on -1 PRF. All the other parameters are assumed to be constant. (a) The effect of  $k_{pas2}$  on the level of frameshift efficiency (FS%, blue line), frameshift protein incorporating a -1 frame aa-tRNA at the recoding site ( $FS_{-1}$ , green line) and frameshift protein incorporating a zero frame aa-tRNA at the recoding site ( $FS_0$ , red line). (b) The effect of  $k_{pas2}$  on the fraction of  $FS_{-1}$ .

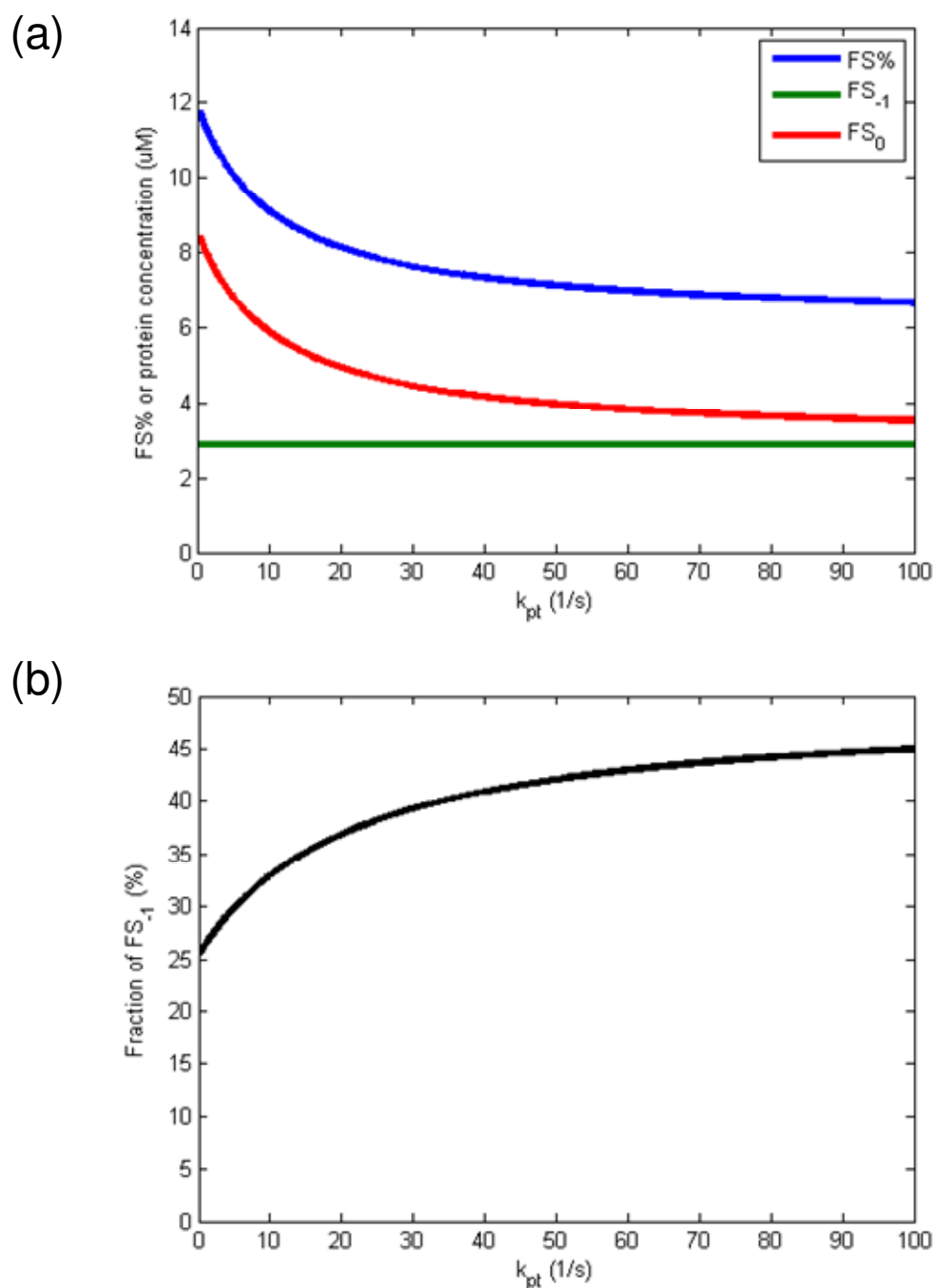


Figure 7.6. The effect of the peptidyl transfer (represented by  $k_{pt}$ ) on -1 PRF. All the other parameters are assumed to be constant. (a) The effect of  $k_{pt}$  on the level of frameshift efficiency (FS%, blue line), frameshift protein incorporating a -1 frame aa-tRNA at the recoding site ( $FS_{-1}$ , green line) and frameshift protein incorporating a zero frame aa-tRNA at the recoding site ( $FS_0$ , red line). (b) The effect of  $k_{pt}$  on the fraction of  $FS_{-1}$ .

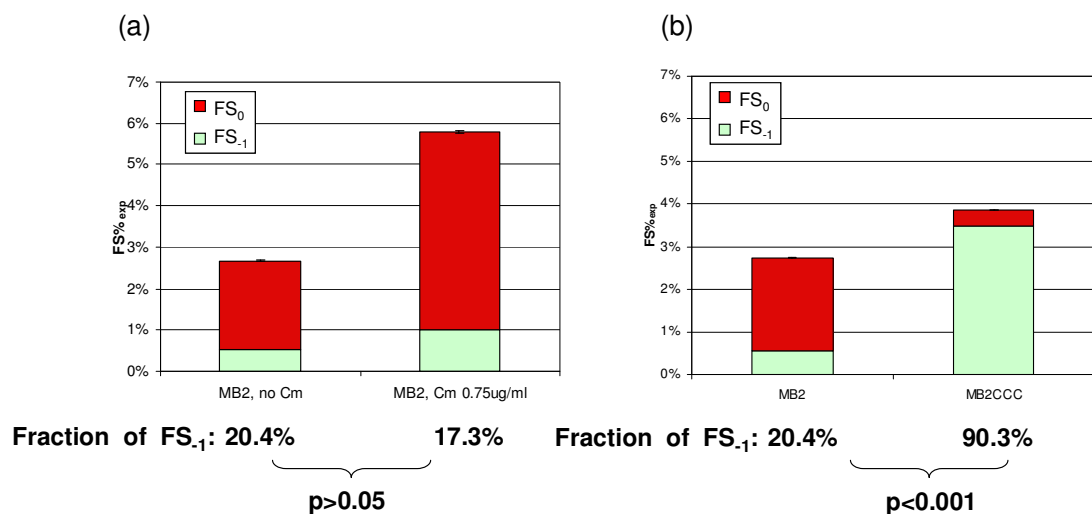


Figure 7.7. Experimentally perturbing the system results in different levels of frameshift efficiency (FS%<sub>exp</sub>) and the fraction of FS<sub>-1</sub>. The total height of the column represents FS%<sub>exp</sub>. The green and the red portions in columns demonstrate the fraction of FS<sub>-1</sub> and FS<sub>0</sub>, respectively. The value for the fraction of FS<sub>-1</sub> is also shown under each column. Error bars indicate the standard deviation for FS%<sub>exp</sub>. (a) The presence of chloramphenicol (Cm) results in higher FS%<sub>exp</sub>. The fraction of the FS<sub>-1</sub> with and without the drug is not significantly different (p>0.05). (b) The MB2CCC strain results in higher FS%<sub>exp</sub> and FS<sub>-1</sub> fraction compared to MB2 (linker sequences listed in Table 7.1).

incomplete translation (Pathway I) involves E- and P-site tRNAs to interact with the -1 frame, CCC as the -1 frame E-site codon may enhance this reaction, *i.e.* an increase in  $r_t$  in the model. The model predicts that a larger  $r_t$  results in a higher FS% and a higher fraction of FS<sub>-1</sub>. Consistently, a 1.5-fold increase in FS%<sub>exp</sub> is observed for the MB2CCC strain compared to the MB2 strain (Figure 7.7.b). In the MB2CCC strain, 90.3% of the frameshift products are FS<sub>-1</sub>. This result suggests that by changing the sequence to favor incomplete translocation, the composition of the frameshift product can be dramatically altered.

## 7.7 Discussion

In this study, a mathematic framework is developed for -1 PRF. To our knowledge, this is the first kinetic model to explain the existence of two types of -1 frameshift proteins. A ribosome can switch the reading frame during incomplete translocation, producing a frameshift product incorporating a -1 frame aa-tRNA in the frameshift site (FS<sub>-1</sub>). Alternatively, a ribosome can switch the reading frame due to a slippage of P- and A-site tRNAs, generating a frameshift product incorporating a zero frame aa-tRNA (FS<sub>0</sub>) in the frameshift site. Previous studies suggested that FS<sub>-1</sub> may result from a single slippage of P-site tRNA [1,24,25]. However, it is not clear that when and how this single slippage of P-site tRNA occurs. In addition, the single slippage model does not explain the experimental evidence regarding the influence of translocation on -1 PRF [21,22]. Our model suggests that both mechanisms, the incomplete translocation and the slippage of P- and A-site tRNAs, participate in making frameshift proteins at various extents for different -1 PRF signals. The frameshifting of the HIV-1 sequence was reported to generate 70% FS<sub>0</sub> and 30% FS<sub>-1</sub> [1,24], which may indicate a stronger influence of the slippage of P-site and A-site tRNAs than the incomplete translocation on FS%. Notably, our protein analysis showed about 80% FS<sub>0</sub> and 20% FS<sub>-1</sub> for the



frameshifting signal in HIV-1 group M subtype B. The discrepancy may be due to the use of different reporter systems or the quantitative methods employed for the assay, which is discussed in Chapter 6. Frameshift products from other -1 PRF signals were analyzed previously [2,33]. For SARS-CoV frameshifting, FS<sub>-1</sub> was not found [2]. For Alphavirus coding sequence 6k, both FS<sub>0</sub> and FS<sub>-1</sub> were found in the frameshift products, but the exact ratio was not determined [33].

In the presence of chloramphenicol, a 2.1-fold increase in FS%<sub>exp</sub> was observed while the fraction of FS<sub>-1</sub> was not significantly different compared to the culture condition without the chemical (Figure 7.7.a). The model predicts that  $k_{pt}$  has a relatively smaller effect on FS% and the fraction of FS<sub>-1</sub> than  $r_t$  and  $k_{pas2}$  (Figure 7.6). A dual fluorescence reporter can sensitively detect small change in FS% in *E. coli* and mammalian cells [26,29,34]. On the other hand, analyzing the composition of the frameshift products relies on steps of protein purification, gel electrophoresis, in gel trypsin digestion, liquid chromatography, and mass spectrometry. The multistage preparation may cause variation in the sample yield, making the detection of a small change difficult.

Mutation of the -1 frame E-site sequence to CCC in the HIV-1 frameshift site may enhance incomplete translocation by allowing more interactions between E-site tRNA and the -1 frame. A significant increase in the fraction of FS<sub>-1</sub> was observed in the MB2CCC strain (Figure 7.7.b). This result is consistent with the model that two mechanisms exist and participate in making frameshift proteins to different extents. The creation of a favorable condition for one pathway can affect the composition of frameshift proteins significantly. To date, no mutations to change the composition of frameshift proteins have been reported in the literature. Notably, our experimental

results show that in one case, FS% increases significantly without a change in the composition of frameshift products (Figure 7.7.a), while in another condition FS% increases a smaller amount but the composition of frameshift products change dramatically (Figure 7.7.b).

## 7.8 Supplementary data

### 7.8.1 Mathematic model

In the kinetic model (Figure 7.2), the formation rate of each component can be written as the following:

$$\frac{d[P_2A_2^{pt}]}{dt} = k_{pt}[P_2A_2] \quad (\text{Eq1})$$

$$\frac{d[P_2A_2]}{dt} = k_7[P_2A_2^*] - k_{pt}[P_2A_2] \quad (\text{Eq2})$$

$$\frac{d[P_2A_2^*]}{dt} = k_5[P_2A_2^{GDP}] - (k_7 + k_6)[P_2A_2^*] \quad (\text{Eq3})$$

$$\frac{d[P_2A_2^{GDP}]}{dt} = k_4[P_2A_2^{GTPase^*}] - k_5[P_2A_2^{GDP}] \quad (\text{Eq4})$$

$$\frac{d[P_2A_2^{GTPase^*}]}{dt} = k_3[E_{02}P_{02}A_2^{CR}] - k_4[P_2A_2^{GTPase^*}] \quad (\text{Eq5})$$

$$\frac{d[E_{02}P_{02}A_2^{CR}]}{dt} = k_2[E_{02}P_{02}A_2^i] - (k_3 + k_{-2})[E_{02}P_{02}A_2^{CR}] \quad (\text{Eq6})$$

$$\frac{d[E_{02}P_{02}A_2^i]}{dt} = k_1[A_2][E_{02}P_{02}] + k_{-2}[E_{02}P_{02}A_2^{CR}] - (k_{-1} + k_2)[E_{02}P_{02}A_2^i] \quad (\text{Eq7})$$

$$\frac{d[E_{02}P_{02}]}{dt} = r_5[E_2P_2^{EFGrl}] + k_{-1}[E_{02}P_{02}A_2^i] + k_{-i}[E_{02}P_{02}A_0^i] - k_1[A_2][E_{02}P_{02}] - k_i[A_0][E_{02}P_{02}]$$

(Eq8)

$$\frac{d[E_{02}P_{02}^{EFGrl}]}{dt} = r_r[E_2P_2^{EFGgdp}] - r_5[E_2P_2^{EFGrl}] \quad (\text{Eq9})$$

$$\frac{d[E_{02}P_{02}^{EFGgdp}]}{dt} = r_t[E_0P_0^{EFGgdp}] - (r_{-t} + r_4)[E_2P_2^{EFGgdp}] \quad (\text{Eq10})$$

$$\frac{d[E_{02}P_{02}A_0^i]}{dt} = k_i[A_0][E_{02}P_{02}] + k_{-i2}[E_{02}P_{02}A_0^{CR}] - (k_{-i} + k_{i2})[E_{02}P_{02}A_0^i] \quad (\text{Eq11})$$

$$\frac{d[E_{02}P_{02}A_0^{CR}]}{dt} = k_{i2}[E_{02}P_{02}A_0^i] - (k_{i3} + k_{-i2})[E_{02}P_{02}A_0^{CR}] \quad (\text{Eq12})$$

$$\frac{d[P_{02}A_0^{GTPase^*}]}{dt} = k_{i3}[E_{02}P_{02}A_0^{CR}] - k_{i4}[P_2A_0^{GTPase^*}] \quad (\text{Eq13})$$

$$\frac{d[P_{02}A_{02}^{pt}]}{dt} = k_{pt}[P_{02}A_{02}] \quad (\text{Eq14})$$

$$\frac{d[P_{02}A_{02}]}{dt} = k_{pas2}[P_0A_0] + k_{s7}[P_{02}A_{02}^*] - (k_{pt} + k_{-pas2})[P_{02}A_{02}] \quad (\text{Eq15})$$

$$\frac{d[P_{02}A_{02}^*]}{dt} = k_{s5}[P_{02}A_{02}^{GDP}] - (k_6 + k_{s7})[P_{02}A_{02}^*] \quad (\text{Eq16})$$

$$\frac{d[P_{02}A_{02}^{GDP}]}{dt} = k_{s4}[P_{02}A_{02}^{GTPase^*}] + k_{i4}[P_{02}A_0^{GTPase^*}] - k_{s5}[P_{02}A_{02}^{GDP}] \quad (\text{Eq17})$$

$$\frac{d[P_{02}A_{02}^{GTPase^*}]}{dt} = k_{pas1}[P_0A_0^{GTPase^*}] - (k_{s4} + k_{-pas1})[P_{02}A_{02}^{GTPase^*}] \quad (\text{Eq18})$$

$$\frac{d[P_0A_0^{pt}]}{dt} = k_{pt}[P_0A_0] \quad (\text{Eq19})$$

$$\frac{d[P_0A_0]}{dt} = k_7[P_0A_0^*] + k_{-pas2}[P_{02}A_{02}] - (k_{pas2} + k_{pt})[P_0A_0] \quad (\text{Eq20})$$

$$\frac{d[P_0A_0^*]}{dt} = k_5[P_0A_0^{GDP}] - (k_6 + k_7)[P_0A_0^*] \quad (\text{Eq21})$$

$$\frac{d[P_0A_0^{GDP}]}{dt} = k_4[P_0A_0^{GTPase^*}] - k_5[P_0A_0^{GDP}] \quad (\text{Eq22})$$

$$\frac{d[P_0A_0^{GTPase^*}]}{dt} = k_3[E_0P_0A_0^{CR} + P_0A_0^{CR}] + k_{-pas1}[P_{02}A_{02}^{GTPase^*}] - (k_4 + k_{pas1})[P_0A_0^{GTPase^*}] \quad (\text{Eq23})$$

$$\frac{d[E_0P_0A_0^{CR}]}{dt} = k_2[E_0P_0A_0^i] - (k_3 + k_{-2})[E_0P_0A_0^{CR}] \quad (\text{Eq24})$$

$$\frac{d[E_0P_0A_0^i]}{dt} = k_1[E_0P_0][A_0] + k_{-2}[E_0P_0A_0^{CR}] - (k_{-1} + k_2)[E_0P_0A_0^i] \quad (\text{Eq25})$$

$$\frac{d[E_0P_0]}{dt} = r_5[E_0P_0^{EFGrl}] + k_{-1}[E_0P_0A_0^i] - k_1[E_0P_0][A_0] \quad (\text{Eq26})$$

$$\frac{d[P_0A_0^{CR}]}{dt} = k_2[P_0A_0^i] - (k_3 + k_{-2})[P_0A_0^{CR}] \quad (\text{Eq27})$$

$$\frac{d[P_0A_0^i]}{dt} = k_1[P_0][A_0] + k_{-2}[P_0A_0^{CR}] - (k_{-1} + k_2)[P_0A_0^i] \quad (\text{Eq28})$$

$$\frac{d[P_0]}{dt} = k_6([P_0A_0^*] + [P_2A_2^*] + [P_{02}A_{02}^*]) + k_{-1}[P_0A_0^i] - k_1[P_0][A_0] \quad (\text{Eq29})$$

$$\frac{d[E_0P_0^{EFGrl}]}{dt} = r_4[E_0P_0^{EFGgdp}] - r_5[E_0P_0^{EFGrl}] \quad (\text{Eq30})$$

$$\frac{d[E_0P_0^{EFGgdp}]}{dt} = r_{TL0}[PA^{EFGgdppi^*}] + r_{-t}[E_{02}P_{02}^{EFGgdp}] - (r_4 + r_t)[E_0P_0^{EFGgdp}] \quad (\text{Eq31})$$

$$\frac{d[PA^{EFGgdppi^*}]}{dt} = r_3[PA^{EFGgdppi}] - r_{TL0}[PA^{EFGgdppi^*}] \quad (\text{Eq32})$$

$$\frac{d[PA^{EFGgdppi}]}{dt} = r_2[PA^{EFGgtp}] - r_3[PA^{EFGgdppi}] \quad (\text{Eq33})$$

$$\frac{d[PA^{EFGgtp}]}{dt} = r_1[PA] - (r_{-1} + r_2)[PA^{EFGgtp}] \quad (\text{Eq34})$$

By assuming steady state, the formation rates of the intermediates equal to zero. The expressions of non-frameshift proteins NFS ( $P_0A_0^{pt}$ ), frameshift proteins FS<sub>-1</sub> incorporating -1 frame aa-tRNAs in the frameshift site ( $P_2A_2^{pt}$ ) and frameshift proteins FS<sub>0</sub> incorporating zero frame aa-tRNAs in the frameshift site ( $P_{02}A_{02}^{pt}$ ) in terms of PA are solved by Matlab v.R2008a (MathWorks Inc., USA). Table 7.S1-7.S4 lists the parameter values used in the model.

Table 7.S1 The rate constants for different steps during translocation at 37°C.

Step	Symbols	Rate constants (s <sup>-1</sup> )
EF-G:GTP binding	$r_1$	150 <sup>a,b</sup>
	$r_{-1}$	140 <sup>a</sup>
EF-G catalyzed GTP hydrolysis	$r_2$	250 <sup>a</sup>
EF-G conformation change	$r_3$	35 <sup>a</sup>
tRNA movement	$r_{TL0}$	500 <sup>c</sup>
Rearrangement of ribosome	$r_4$	5 <sup>a</sup>
EF-G dissociation	$r_5$	20 <sup>a</sup>
	$r_t$	1 <sup>d</sup>
	$r_{-t}$	10 <sup>d</sup>

<sup>a</sup> Wintermeyer *et al.* [35]

<sup>b</sup>  $\mu\text{M}^{-1}\text{s}^{-1}$

<sup>c</sup> Assumed value in the study. A rapid reaction according to Wintermeyer *et al.* [35]

<sup>d</sup> Assume value in the study

Table 7.S2 The rate constants for different steps in aminoacyl-tRNA selection at 20°C. The rate constants used in the model equal the rate constant at 20°C multiplied by fold change from 20°C to 37°C (Table 7.S3).

Step	Symbols	Rate constants (s <sup>-1</sup> )
Initial binding	k <sub>1</sub> , k <sub>i</sub>	110 <sup>a,b</sup>
	k <sub>-1</sub> , k <sub>-i</sub>	25 <sup>a</sup>
Codon recognition	k <sub>2</sub> , k <sub>i2</sub>	100 <sup>a</sup>
	k <sub>-2</sub> , k <sub>-i2</sub>	0.2 <sup>a</sup>
GTPase activation and GTP hydrolysis*	k <sub>3</sub> , k <sub>4</sub> , k <sub>s4</sub>	260 <sup>a</sup>
EF-Tu conformational change (dissociation)	k <sub>5</sub> , k <sub>s5</sub>	60 <sup>a</sup>
tRNA rejection	k <sub>6</sub>	0.3 <sup>a</sup>
Accommodation	k <sub>7</sub> , k <sub>s7</sub>	7 <sup>a</sup>
	k <sub>pt</sub>	50 <sup>c</sup>
Simultaneous slippage before GTP hydrolysis	k <sub>pas1</sub>	1 <sup>d</sup>
	k <sub>-pas1</sub>	10 <sup>d</sup>
Simultaneous slippage before peptidyl transfer	k <sub>pas2</sub>	1 <sup>d</sup>
	k <sub>-pas2</sub>	10 <sup>d</sup>

<sup>a</sup> Rodnina *et al.* [36].

<sup>b</sup> μM<sup>-1</sup>s<sup>-1</sup>

<sup>c</sup> Rate constant at 37°C; Katunin *et al.* [37]

<sup>d</sup> Assumed values in the model

Table 7.S3 The activation energy for different steps in the model and the fold change of the rate constants ( $k_{310K}/k_{293K}$ , from 20°C to 37°C).

	$E_a$ (kJ/mol)	$k_{310K}/k_{293K}$
$k_1, k_{1s}$	$10 \pm 6^a$	1.25
$k_{-1}, k_{-1s}$	$46 \pm 5^a$	2.82
$k_2, k_{2s}$	$38 \pm 8^a$	2.36
$k_{-2}, k_{-2s}$	$44 \pm 5^a$	2.7
$k_3, k_{3s}$	$55^b$	3.45
$k_4, k_{4s}$	$55^b$	3.45
$k_5, k_{5s}$	$155^d$	33
$k_6, k_{6s}$	$55^e$	3.45
$k_7, k_{7s}$	$55^e$	3.45

<sup>a</sup> Rodnina *et al.* [38].

<sup>b</sup> Thompson *et al.* [39].

<sup>c</sup> Gromadski *et al.* [40].

<sup>d</sup> Karim *et al.* [41]

<sup>e</sup> Assumed values in the model

Table 7.S4 The concentration of the components used in the model.

Components	Symbols	Parameter values
Initial reactant	PA	1 <sup>a</sup>
Zero-frame aa-tRNA	$A_0$ (%)	10 <sup>a</sup>
-1 frame aa-tRNA	$A_2$ (%)	10 <sup>a</sup>

<sup>a</sup> Assumed value in the model ( $\mu\text{M}$ ).

Table 7.S5 The list of target peptides and their MRM parameters. All peptides were confirmed by a Mascot search of their MS/MS spectra against a custom-made database ( $p < 0.05$ ).

Strain	FS <sup>a</sup>	Sequence	Theor MW <sup>b</sup>	Prec $m/z$ <sup>c</sup>	Frag $m/z$ <sup>d</sup>	$z$ <sup>e</sup>	CE <sup>f</sup>
MB2/MB2Cm	FS <sub>0</sub>	HSTGAASTANFLR	1331.68	444.90	620.40	3	22.580
MB2/MB2Cm	FS <sub>-1</sub>	HSTGAASTANFFR	1365.58	456.20	654.30	3	23.073
MB2CCC	FS <sub>0</sub>	HSTGAASTAPFLR	1314.67	439.23	532.32	3	22.326
MB2CCC	FS <sub>-1</sub>	HSTGAASTAPFFR	1348.63	450.55	566.30	3	22.824

<sup>a</sup> Type of frameshift products

<sup>b</sup> Theoretical molecular weight (Da)

<sup>c</sup> Precursor ion  $m/z$

<sup>d</sup> Fragment ion  $m/z$

<sup>e</sup> Charge state

<sup>f</sup> Collision energy (V)

## 7.8.2 Sensitivity analysis

A program was developed in Matlab v.R2008a to perform an n-way analysis of variance (ANOVA). Each parameter can vary for 5 levels: a base line value,  $\pm 25\%$  of the base line,  $\pm 50\%$  of the base line. Randomly selected 10,000 parameter sets are used to calculate FS%. The analysis calculates F statistic value for each tested parameter (Figure 7.S1). A higher F statistic indicates a larger impact of the parameter on FS%.

## 7.9 Conclusion

A mathematic framework is developed for -1 PRF upon the translation elongation cycle. The model presents not only the change in frameshift efficiency, but also the change in the composition of frameshift products under different conditions. In addition, the model identifies dominating parameters, representing steps in the translation elongation, on -1 frameshifting. Experimentally targeting these steps results in different levels of frameshift efficiency, which are consistent with model predictions. A mutation in the -1 frame E-site sequence was shown to dramatically



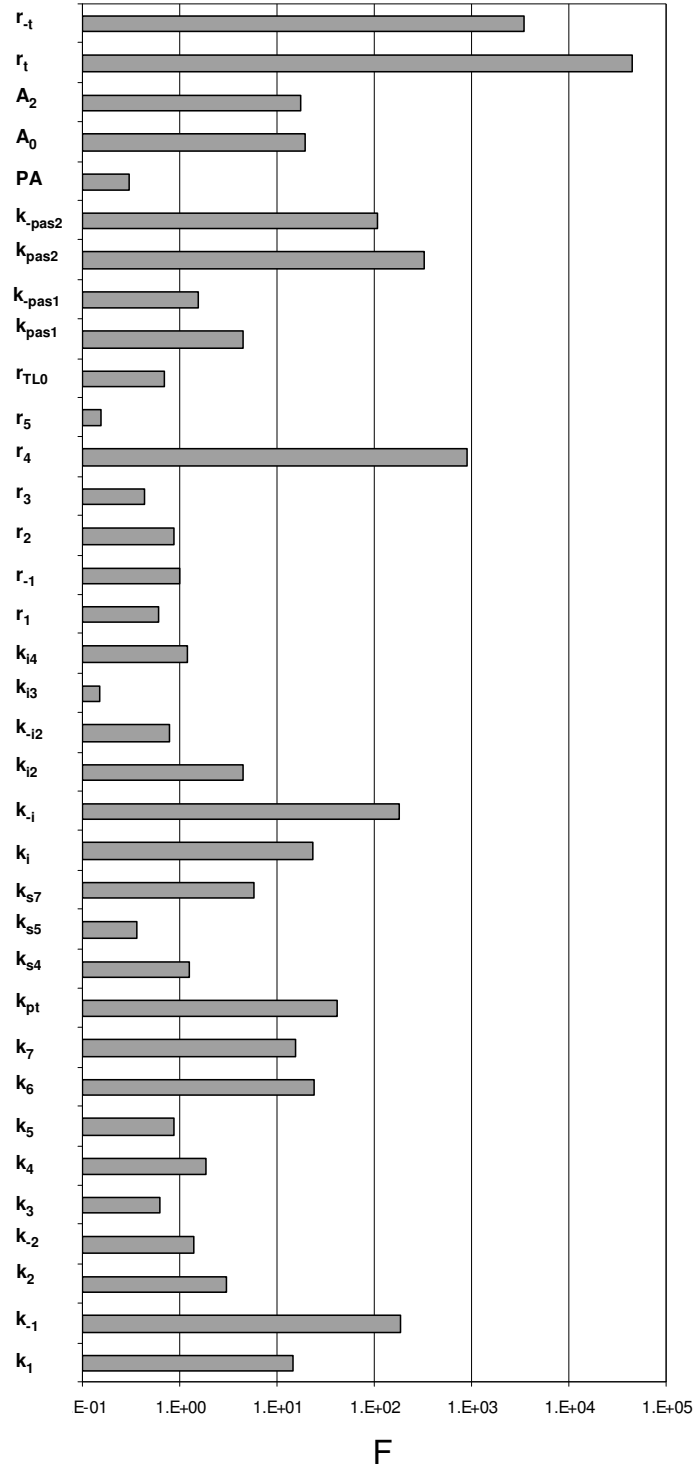


Figure 7.S1. Sensitivity analysis using n-way analysis of variance (ANOVA). A higher F statistic value suggests a more significant impact of the parameter on FS%.

change the composition of frameshift products, suggesting an important role for the sequence upstream of the slippery site. Our results suggest that not only the frameshift efficiency, but also the compositions of the frameshift products are worth investigated to advance our knowledge for -1 PRF.

## REFERENCES

1. Jacks,T., Power,M.D., Masiarz,F.R., Luciw,P.A., Barr,P.J. and Varmus,H.E. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280-283.
2. Baranov,P.V., Henderson,C.M., Anderson,C.B., Gesteland,R.F., Atkins,J.F. and Howard,M.T. (2005) Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology*, **332**, 498-510.
3. Biswas,P., Jiang,X., Pacchia,A.L., Dougherty,J.P. and Peltz,S.W. (2004) The human immunodeficiency virus type 1 ribosomal frameshifting site is an invariant sequence determinant and an important target for antiviral therapy. *J. Virol.*, **78**, 2082-2087.
4. Plant,E.P. and Dinman,J.D. (2008) The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front. Biosci.*, **13**, 4873-4881.
5. Dinman,J.D., Ruiz-Echevarria,M.J. and Peltz,S.W. (1998) Translating old drugs into new treatments: Ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.*, **16**, 190-196.
6. Brierley,I., Jenner,A.J. and Inglis,S.C. (1992) Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.*, **227**, 463-479.
7. Jacks,T., Madhani,H.D., Masiarz,F.R. and Varmus,H.E. (1988) Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. *Cell*, **55**, 447-458.
8. Brierley,I., Digard,P. and Inglis,S.C. (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell*, **57**, 537-547.
9. ten Dam,E.B., Pleij,C.W. and Bosch,L. (1990) RNA pseudoknots: Translational frameshifting and readthrough on viral RNAs. *Virus Genes*, **4**, 121-136.
10. Tu,C., Tzeng,T.H. and Bruenn,J.A. (1992) Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 8636-8640.
11. Somogyi,P., Jenner,A.J., Brierley,I. and Inglis,S.C. (1993) Ribosomal pausing during translation of an RNA pseudoknot. *Mol. Cell. Biol.*, **13**, 6931-6940.
12. Lopinski,J.D., Dinman,J.D. and Bruenn,J.A. (2000) Kinetics of ribosomal pausing during programmed -1 translational frameshifting. *Mol. Cell. Biol.*, **20**, 1095-1103.

13. Kollmus,H., Honigman,A., Panet,A. and Hauser,H. (1994) The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II *in vivo*. *J. Virol.*, **68**, 6087-6091.
14. Farabaugh,P.J. (1997). Programmed alternative reading of the genetic code. R. G. Landes Co., Austin, TX, pp. 69-102.
15. Plant,E.P., Jacobs,K.L., Harger,J.W., Meskauskas,A., Jacobs,J.L., Baxter,J.L., Petrov,A.N. and Dinman,J.D. (2003) The 9-A solution: How mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. *RNA*, **9**, 168-174.
16. Noller,H.F., Yusupov,M.M., Yusupova,G.Z., Baucom,A. and Cate,J.H. (2002) Translocation of tRNA during protein synthesis. *FEBS Lett.*, **514**, 11-16.
17. Dinman,J.D. and Kinzy,T.G. (1997) Translational misreading: Mutations in translation elongation factor 1alpha differentially affect programmed ribosomal frameshifting and drug sensitivity. *RNA*, **3**, 870-881.
18. Harger,J.W., Meskauskas,A. and Dinman,J.D. (2002) An "integrated model" of programmed ribosomal frameshifting. *Trends Biochem. Sci.*, **27**, 448-454.
19. Leger,M., Sidani,S. and Brakier-Gingras,L. (2004) A reassessment of the response of the bacterial ribosome to the frameshift stimulatory signal of the human immunodeficiency virus type 1. *RNA*, **10**, 1225-1235.
20. Kim,Y.G., Maas,S. and Rich,A. (2001) Comparative mutational analysis of cis-acting RNA signals for translational frameshifting in HIV-1 and HTLV-2. *Nucleic Acids Res.*, **29**, 1125-1131.
21. Leger,M., Dulude,D., Steinberg,S.V. and Brakier-Gingras,L. (2007) The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed -1 ribosomal frameshift. *Nucleic Acids Res.*, **35**, 5581-5592.
22. Weiss,R.B., Dunn,D.M., Shuh,M., Atkins,J.F. and Gesteland,R.F. (1989) *E. coli* ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. *New Biol.*, **1**, 159-169.
23. Namy,O., Moran,S.J., Stuart,D.I., Gilbert,R.J. and Brierley,I. (2006) A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature*, **441**, 244-247.
24. Yelverton,E., Lindsley,D., Yamauchi,P. and Gallant,J.A. (1994) The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in *Escherichia coli*. *Mol. Microbiol.*, **11**, 303-313.

25. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2004) P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221-230.
26. Liao,P.Y., Gupta,P., Petrov,A.N., Dinman,J.D. and Lee,K.H. (2008) A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Res.*, **36**, 2619-2629.
27. Savelsbergh,A., Katunin,V.I., Mohr,D., Peske,F., Rodnina,M.V. and Wintermeyer,W. (2003) An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell*, **11**, 1517-1523.
28. Rodnina,M.V., Gromadski,K.B., Kothe,U. and Wieden,H.J. (2005) Recognition and selection of tRNA in translation. *FEBS Lett.*, **579**, 938-942.
29. Liao,P.Y., Choi,Y.S. and Lee,K.H. (2009) FSscan: A mechanism-based program to identify +1 ribosomal frameshift hotspots. *Nucleic Acids Res.*, doi:10.1093/nar/gkp796.
30. Jacobs,J.L. and Dinman,J.D. (2004) Systematic analysis of bicistronic reporter assay data. *Nucleic Acids Res.*, **32**, e160.
31. Dinman,J.D., Ruiz-Echevarria,M.J., Czaplinski,K. and Peltz,S.W. (1997) Peptidyl-transferase inhibitors have antiviral properties by altering programmed -1 ribosomal frameshifting efficiencies: Development of model systems. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 6606-6611.
32. Schlunzen,F., Zarivach,R., Harms,J., Bashan,A., Tocilj,A., Albrecht,R., Yonath,A. and Franceschi,F. (2001) Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature*, **413**, 814-821.
33. Firth,A.E., Chung,B.Y., Fleeton,M.N. and Atkins,J.F. (2008) Discovery of frameshifting in alphavirus 6K resolves a 20-year enigma. *Viol. J.*, **5**, 108.
34. Cardno,T.S., Poole,E.S., Mathew,S.F., Graves,R. and Tate,W.P. (2009) A homogeneous cell-based bicistronic fluorescence assay for high-throughput identification of drugs that perturb viral gene recoding and read-through of nonsense stop codons. *RNA*, **15**, 1614-1621
35. Wintermeyer W., Peske F., Beringer M., Gromadski K.B., Savelsbergh A. and Rodina M.V. (2004) Mechanisms of elongation on the ribosome: dynamics of a macromolecular machine. *Biochemical society Transactions*, **22**, 733-737.
36. Rodnina,M.V., Gromadski,K.B., Kothe,U., Wieden,H.J. (2005) Recognition and selection of tRNA in translation. *FEBS Lett.*, **579**, 938-942.

37. Katunin, V.I., Muth, G.W., Strobel, S.A., Wintermeyer, W. and Rodnina, M.V. (2002) Important contribution to catalysis of peptide bond formation by a single ionizing group within the ribosome. *Mol. Cell*, **10**, 339-346.
38. Rodnina, M.V., Pape, T., Fricke, R., Kuhn, L., Wintermeyer, W. (1996) Initial binding of the elongation factor tu.GTP.aminoacyl-tRNA complex preceding codon recognition on the ribosome. *J. Biol. Chem.*, **271**, 646-652.
39. Thompson, R.C., Dix, D.B., Eccleston, J.F. (1980) Single turnover kinetic studies of guanosine triphosphate hydrolysis and peptide formation in the elongation factor tu-dependent binding of aminoacyl-tRNA to *Escherichia coli* ribosomes. *J. Biol. Chem.*, **255**: 11088-11090.
40. Gromadski, K.B., Daviter, T. and Rodnina, M.V. (2006) A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Mol. Cell* **21**, 369-377.
41. Karim, A.M. and Thompson, R.C. (1986) Guanosine 5'-O-(3-thiotriphosphate) as an analog of GTP in protein biosynthesis. The effects of temperature and polycations on the accuracy of initial recognition of aminoacyl-tRNA ternary complexes by ribosomes. *J. Biol. Chem.*, **261**, 3238-3243.

## CHAPTER 8

### FROM SNPS TO FUNCTIONAL POLYMORPHISM: THE INSIGHT INTO BIOTECHNOLOGY APPLICATIONS

#### ***8.1 Preface***

During my Ph.D studies, I was given an opportunity to learn and write about a different aspect of molecular biology that can impact protein expression. A single nucleotide polymorphism (SNP) is the most common genetic variation in the human genome. SNP studies provide opportunities to understand how nucleotide variations contribute to altered gene expressions, which may result in different phenotypes. This chapter is a review for SNP analysis and its potential for biotechnology applications.

#### ***8.2 Abstract***

Single nucleotide polymorphisms (SNP) are the most common form of genetic variation in the genome. Scanning a genome for SNPs can help identify millions of potentially informative biomarkers. SNPs have been extensively used as molecular markers in human disease genetics, pharmacogenetics, and breeding, but SNPs have not been widely used in the bioprocess community. In biotechnology applications such as bioprocess development, SNPs may serve as genetic markers for phenotypes of interest such as those related to cell growth and viability, specific productivity, or stability. Furthermore, SNPs that relate to particular phenotypes may be targets for metabolic and cellular engineering. This review introduces study designs that have been used to link SNPs and phenotypes. The review then focuses on the downstream effects of the SNPs at DNA, RNA and protein levels. Finally, this review discusses specific examples to apply SNPs for breeding, strain evolution, and biomolecule production. Large scale SNP studies represent an opportunity to apply new genome-

scale technologies to address current limitations and questions relevant to the biotechnology community such as cell line generation and selection.

### **8.3 Introduction**

A single nucleotide polymorphism (SNP) is a nucleotide variation at a specific location in the genome (Figure 8.1). (Table 8.1 provides a glossary of terms in SNP studies). Typically, SNPs are bi-allelic and by definition found in more than 1% of the population [1]. In practice, tri- or tetra-allelic SNPs, insertions, deletions and variations found in less than 1% of the population are also referred as SNPs. SNPs are the most abundant variations in the human genome [2,3]. The International HapMap Project has characterized over 3.1 million human SNPs, indicating a SNP density of approximately one per kilobase [4].

Discrimination of genetic variants has great potential to unravel the molecular basis for “super producer strains” and to gain insight into functional genomics for biotechnology applications. Although genotyping of common recombinant hosts (e.g. Chinese hamster ovary or CHO cells) has not been reported in the literature, SNP analysis has been widely used in human disease genetics, pharmacogenetics, and breeding. These studies attempt to predict diseases, drug responses, and breeds with higher economic value based on the variations in the genetic sequences. We believe that these same approaches may help the bioprocess research community predict strain performance by searching for specific SNPs in relevant recombinant host cells. Furthermore, an understanding of the downstream effects of the genetic variation is critically important and can provide targets for metabolic and cellular engineering studies to design hosts with specific phenotypic characteristics of interest to the bioprocess community. For example, newly introduced transgene sequences can be



Subject 1	TCGACT <b>A</b> CTCTA...CGTT <b>C</b> AGGCGT...AC <b>G</b> CATTACGGCGTCC
Subject 2	TCGACT <b>G</b> CTCTA...CGTT <b>T</b> AGGCGT...AC <b>A</b> CATTAGGGCGTCC
Subject 3	TCGACT <b>A</b> CTCTA...CGTT <b>C</b> AGGCGT...AC <b>A</b> CATTACGGCGTCC
Subject 4	TCGACT <b>G</b> CTCTA...CGTT <b>C</b> AGGCGT...AC <b>G</b> CATTACGGCGTCC
Subject 5	TCGACT <b>A</b> CTCTA...CGTT <b>C</b> AGGCGT...AC <b>A</b> CATTATGGCGTCC
Subject 6	TCGACT <b>A</b> CTCTA...CGTT <b>C</b> AGGCGT...AC <b>A</b> CATTACGGCGTCC

<b>SNP</b>	<b>A/G</b>	<b>C/T</b>	<b>G/A</b>	<b>C/G/T</b>
<b>Haplotype</b>		A-C-G-C		
		G-T-A-G		
		A-C-A-C		
		G-C-G-C		
		A-C-A-T		

Figure 8.1. An example of single nucleotide polymorphisms (SNPs) and the haplotype observed in six subjects (concept adapted from [2]). Four SNPs (A/G, C/T, G/A, and C/G/T) are found, where A/G, C/T and G/A are bi-allelic SNPs and C/G/T is a tri-allelic SNP. Theoretically, these four SNPs allow 24 haplotypes, but only five haplotypes are found in the six subjects.

Table 8.1 Glossary of terms in single nucleotide polymorphism (SNP) studies

Term	Description
Alleles	Alternate forms of a gene of chromosomal locus that differ in DNA sequence
Biallelic	Only two of the four common nucleotides are found in a specific position
Expressed sequence tag	Short sub-sequence of a transcribed cDNA sequence that may be used for gene identification
Genotype	Inheritable genetic constitution carried by living organisms
Haplotype	A set of alleles located at neighboring genes or genomic sequences that tend to be inherited together
HapMap	Genome-wide database of common genetic sequence variation in human
Non-synonymous SNP	A single nucleotide variation in the coding sequence that results in a change in amino acid sequence
Pharmacogenomics	The study to understand the effect of genetic polymorphisms on drug responses
Phenotype	The physical manifestation of genetic information
Single nucleotide polymorphism (SNP)	A single nucleotide variation in the genetic sequence
Synonymous SNP	A single nucleotide variation in the coding sequence that results in no change in amino acid sequence

adjusted to ensure optimal production or host metabolism can be modified to favor the recombinant protein expression. Figure 8.2 shows a proposed framework from SNP identification to potential bioprocess applications. Following this framework, this review first introduces how SNPs are discovered and associated to a phenotype of interest. The second part describes different mechanisms of how SNPs can contribute to phenotypes. Finally, several biotechnology applications will be discussed.

#### **8.4 *SNP analysis***

The recent availability of high-throughput genotyping technologies has made large scale genome scans possible. Several reviews are available for genotyping technologies [5-8] and our goal here is not to present an extremely detailed treatment of all available technologies; we direct the reader to those reviews for such an analysis. In general, SNP analysis involves three steps. First, SNPs are identified and mapped onto known gene sequences or genomes. Second, the genome sequence can be scanned for the presence or absence of known SNPs. Third, SNPs in the genome are linked to a specific phenotype.

##### **8.4.1 SNP discovery**

Several methods are used for SNP discovery. One of the most straightforward ways to discover novel SNPs is to sequence DNA fragments amplified by polymer chain reaction (PCR). PCR primers are designed to amplify both strands of DNA from genes or other single copy genomic sequences. PCR products are sequenced and aligned into gene sequences, allowing novel SNP identifications. The International HapMap project has primarily used this approach to discover SNPs in the human genome [3]. SNP discovery by sequencing amplified DNA fragments is very reliable, with the false discovery rate below 5% [6]. However, this method is costly and requires enormous

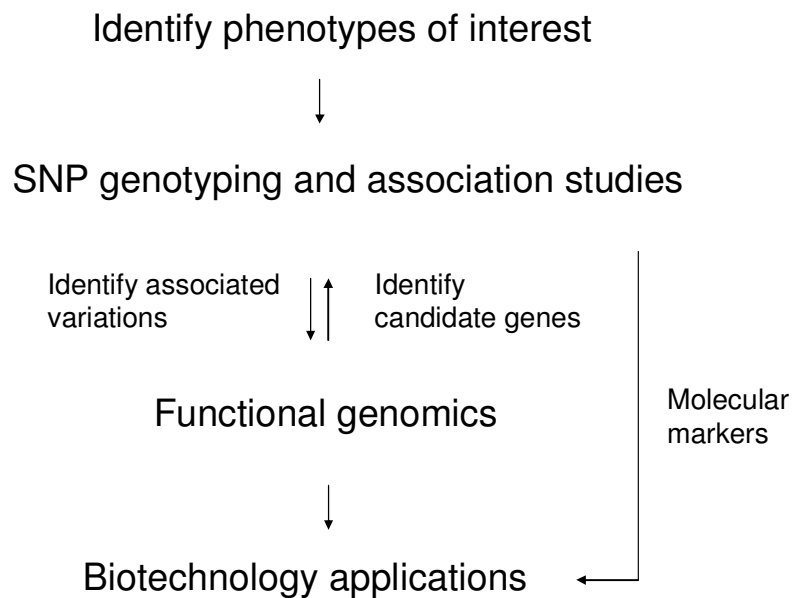


Figure 8.2. A possible framework for SNP applications in the biotechnology industry. The approach starts with identifying phenotypes of interest. The genetic sequences of different subjects are tested for variations. Relevant SNPs are investigated further for their functionalities. The knowledge of functional genomics can provide insight into genetic engineering for biotechnology applications. In addition, functional genomics also help researchers to identify candidate genes for SNP analysis. The molecular markers associated with a specific phenotype can also be directly used as selective markers for strain developments in biotechnology.

effort because specific primers have to be developed and large numbers of sequences need to be amplified and sequenced.

SNP identification can be based on expressed sequence tags (ESTs), which are generated by single-run sequencing of cDNAs. The resulting sequences are relatively short and low quality fragments. Because the majority of EST libraries have been obtained from different individuals, assembly of overlapping sequences for the same region permits novel SNP discoveries. As a result, EST data provides a valuable resource for SNP mining, especially when a reference genome is not available [9-11]. However, because many sequence variations may result from poor quality sequencing, this method has higher false discovery rate from 15 to 50% [6].

Recent advances in so-called next generation sequencing technology enables millions of sequence reads in parallel, allowing whole genome sequencing in a fast and a low cost manner [12]. With a reference genome, re-sequencing using this technology provides a high throughput platform for the SNP discovery [8]. If no reference genome is available, next generation sequencing technology can be applied to sequence ESTs to allow an efficient SNP discovery, although this approach is limited to discovering SNPs in known expressed genes.

#### **8.4.2 SNP detection**

Once a large number of SNPs has been identified, it is critical to quantify the presence of the variation in a larger scale study. Specifically, the gene sequences, or genomes, of different subjects are tested for a set of known SNPs. Kim *et al.* summarized detection methods, analysis scales, strengths and applications for different genotyping technologies based on selected SNPs [7] and the reader is referred to that article for a

detailed treatment. Medium- to high-throughput genotyping technologies, such as BeadArray<sup>TM</sup> (Illumina, San Diego, CA) or MassEXTEND<sup>TM</sup> (Sequenom, San Diego, CA), can assay several hundred SNPs simultaneously [7]. Genechip<sup>®</sup> (Affymetrix, Santa Clara, CA), a microarray based genotyping technology, can assay  $10^4$  to  $10^5$  SNPs throughout the genome [7]. Because genotyping technologies usually involve primer extension near a SNP site or hybridization of sequences spanning a SNP site, the major challenge for these technologies is the requirement of prior knowledge of SNPs. An additional challenge is that most of the technologies available for SNP detection can detect only binary variations. Interestingly, next generation sequencing technology allows SNP discovery coupling with SNP detection and is not limited to binary variations. Although this approach is mostly used in SNP discovery to date, further technology development may lead to a next generation SNP analysis method if appropriate bioinformatic tools become available.

#### **8.4.3 Study designs to relate a SNP to a phenotype**

Two approaches are commonly used to link SNPs to a phenotype of interest: 1) a candidate gene association study and 2) a genome-wide association study (Table 8.2). Although these approaches are developed for human disease genetics, they are applicable to strain characterizations in the biotechnology industry.

A candidate gene association study begins with genes that are suspected to associate with a phenotype of interest [13]. These candidate genes are usually identified from specific biological pathways based on a literature search. This type of the study compares candidate genes in case and control populations. The number of genes under investigation could be 1 to 2 (single-gene based), 10 to 20 (gene family based), or 50 and up (biological pathway based). Because of a small to medium scale analysis, both

the direct sequencing of candidate genes and the use of a genotyping technology to detect known SNPs are applicable for candidate gene association studies. Among these two methods, direct sequencing of candidate genes may lead to novel SNP discoveries. With the completion of the sequencing and annotation of the human genome, where about 45,000 genes have been identified and over twenty-five million SNPs reported in National Center for Biotechnology Information's website, a candidate gene association study becomes an effective approach for identifying phenotype-causing variations in humans. One major challenge for candidate gene association studies is their dependence on prior knowledge about genes or biological pathways, which may be incomplete. Thus, candidate gene association studies are less likely to identify variations in a new gene associated with a phenotype.

A genome-wide association (GWA) study assays more than 100,000 SNPs for hundreds of unrelated subjects. The relationship between a specific genotype and a phenotype is used to characterize susceptibility genes that are associated with observable traits. Because of a large scale analysis, GWA studies use high-throughput genotyping technologies based on selected SNPs [7]. GWA studies provide the opportunity to discover novel genes involved in a phenotype because they do not depend upon prior knowledge of the biological pathways. However, the GWA approach has a potential for false-positive or false-negative results related to the selection of study participants and genotyping errors [14].

### ***8.5 How SNPs lead to different phenotypes***

SNPs are not evenly distributed across the genome. In general, SNPs occur much less frequently in coding regions of the genome than in noncoding regions [15]. SNPs in regulatory sites of a gene can affect transcription rates, thereby changing the

Table 8.2 Different types of genetic studies to connect SNPs to a phenotype

Method	Description
Candidate gene studies	<p>Variations in genes suspected to associate with a phenotype are compared.</p> <p><i>Pros.</i> Increased statistical efficiency of association analysis of polygenic phenotype</p> <p><i>Cons.</i> May be based on imperfect understanding of the biologic pathways</p>
Genome-wide Association (GWA) study	<p>SNPs associated with observable traits across given genomes are identified by using high-throughput genotyping technologies.</p> <p><i>Pros.</i> Search the entire genome for associations without assuming candidates.</p> <p><i>Cons.</i> Has potential for false-positive and false-negative results and for biases related to selection of study participants and genotyping errors</p>



expression of corresponding proteins. In coding regions, exonic SNPs can be categorized into two classes: non-synonymous SNPs that alter the amino acid sequence of the protein products and synonymous SNPs that do not affect primary sequence of the products. Non-synonymous polymorphisms have been more widely characterized because their effects are relatively easy to detect computationally and experimentally. Proteins with the same sequence derived from synonymous SNPs were previously assumed to exert no discernible effect on a gene function or a phenotype. These gene variants are often termed “silent mutations”. However, several synonymous mutations have been reported to alter gene expression or protein folding [16-18]. Recently, a synonymous codon library has been shown to result in different levels of gene expression [19]. These reports demonstrate that synonymous SNPs can also produce different phenotypes.

The molecular effects of SNPs are now better understood in many cases. Several bioinformatic tools have been developed to predict functional SNPs [20]. Based on the central dogma, SNPs should affect phenotypes at the DNA, RNA and protein levels. The following section discusses several mechanisms for how SNPs affect phenotypes. It is important to note that these mechanisms are not exclusive. For example, a non-synonymous SNP may affect gene expression at both RNA and protein levels.

#### **8.5.1 DNA level: from DNA to RNA**

Regulatory polymorphisms can potentially cause variations in gene expression. An early study observed that about a third of the promoter variants in the human genome may alter gene expression by 50% or more [21]. Regulatory polymorphisms can be classified into two groups: a *cis*-acting polymorphism affects genes in or near the locus and a *trans*-acting polymorphism in one gene affects the expression of another

gene at a different locus.

A SNP in a regulatory DNA binding site may alter the affinity with the regulatory protein, resulting in different gene expressions (Figure 8.3). SNPs in the osteopontin promoter have been shown to modify DNA binding affinity to transcription factors SP1/SP3 [22]. A GWA study revealed a G-to-A substitution in the 5' untranslated region (5'-UTR) of the *FOXE1* gene to associate with thyroid cancer susceptibility [23]. Recently, this variant was found to alter the recruitment of USF1/USF2 transcription factors [24]. The T-to-C substitution located in the 5'-UTR of the *GDF5* gene causes a different interaction with DEAF-1, a *trans*-acting factor for *GDF5*, leading to a reduced gene expression [25]. In addition, a SNP in 3'-UTR of *GDF5* can alter the gene expression independent of the SNP in 5'UTR, highlighting the complexity of this gene regulation.

### **8.5.2 RNA level: from RNA to protein**

SNPs can alter mRNA folding and thus affect mRNA stability (Figure 8.4.a). In the dopamine receptor *D2* gene, a synonymous variation C957T resulted in different mRNA structures [26]. The authors suggested that different folding caused the mRNA carrying C957T to be degraded more rapidly than the wild type sequence. As a result, less of the encoded protein is made, leading to cognitive disorders. An intragenic SNP in the *CDSN* gene (CDSN\*971T) decreased the transcript affinity for a 39 kDa RNA binding protein, which increased mRNA stability two-fold and up-regulated the *CDSN* product, corneodesmosin, in patients with Psoriasis, a chronic skin disorder with multifactorial etiology [27]. Similarly, the characterization of two nonsynonymous SNPs, C74A and G223A, in the *mTPH2* gene revealed that A-A, C-A and A-G haplotypes increased mRNA stability and enzyme activity as compared to wild-type

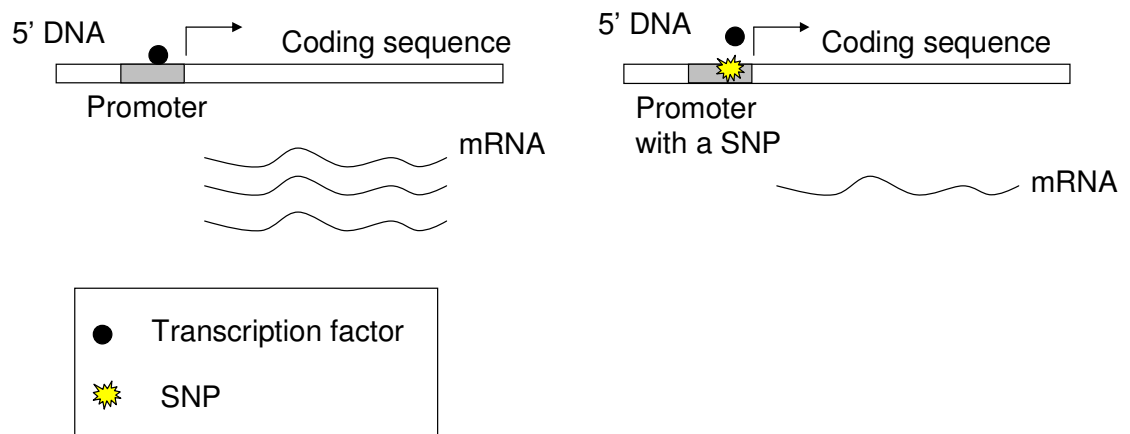


Figure 8.3. An example of the effect of a SNP on DNA and transcriptional levels. Regulatory SNPs may alter DNA affinity to a transcription factor, resulting in different mRNA levels.

C-G haplotype [28]. Because mTPH2 synthesizes neuronal serotonin, the up-regulation of the protein activity is hypothesized to lead to psychiatric disorders.

The different structures of mRNA caused by SNPs may also alter the protein synthesis rate (Figure 8.4.b). In catechol-O-methyltransferase, three haplotypes result in different mRNA local stem-loop structures. The most stable structure is correlated with the lowest protein levels and enzymatic activity [29]. Because catechol-O-methyltransferase is a key regulator of pain perception [30], variations in enzyme activities cause changes in pain sensitivity in patients.

SNPs can affect the efficiency of translation initiation (Figure 8.4.c). In the Kozak sequence of the *hCD40* gene, C-T, T-T and C-C haplotypes were shown to result in similar mRNA levels but different protein levels. The authors suggested that the C polymorphism in the Kozak sequence allowed the ribosome to initiate translation more efficiently, resulting in a higher protein level [31].

By modifying translation elongation, SNPs may alter protein conformations (Figure 8.4.d). Synonymous codon substitutions may lead to different kinetics of protein translation, thus yielding a protein with a different final conformation and function [32]. Kimchi-Sarfaty *et al.* reported that a naturally occurring synonymous SNP can result in altered drug interactions of P-glycoprotein (P-gp) [33]. The authors suggested that the synonymous polymorphism affects the timing of cotranslational folding and the insertion of P-gp into the membrane, thereby altering the structure of substrate and inhibitor interaction sites. This result indicates that synonymous SNPs might contribute to development and progression of certain diseases and should not be neglected.

Pre-mRNA splicing is a complex mechanism that relies on the correct recognition of protein coding sequences (exons) from the non coding sequences (introns) on RNA transcripts [34]. Figure 8.4.e illustrates how SNPs might affect mRNA splicing. Natural mutations D565G and G576A and several site-directed silent substitutions in the cystic fibrosis transmembrane-conductance receptor (CFTR) exon 12 were shown to induce a variable extent of exon skipping [35]. Skipping of this exon removes a part of the first nucleotide-binding domain of CFTR, rendering the protein non-functional.

MicroRNAs (miRNAs) are a class of noncoding RNAs that can regulate gene expression by base pairing with target mRNAs at the 3'-UTRs, leading to an mRNA cleavage or translational repression [36]. SNPs in miRNA genes may alter miRNA processing while SNPs around the miRNA binding sites in the target mRNAs may affect miRNA function. SNPs in miRNA-125a and miRNA-K5 were reported to impair miRNA processing [37]. Sun *et al.* tested 24 human X-linked miRNA variants in schizophrenia and autism and reported that SNPs in miRNA genes can impair or enhance miRNA processing as well as alter the sites of processing [38].

### **8.5.3 Protein level: from polypeptide formation to post-translation modification**

At the protein and post-translational levels, a substantial effort has been invested in the function of non-synonymous SNPs because their downstream effects are relatively easy to characterize. Several reviews have described how non-synonymous SNPs affect protein functions and interactions [39-41]. The following section presents the change in protein activities in terms of protein stabilities, binding affinities, catalytic properties, and post-translational modifications (Figure 8.5).

Variation in protein stability due to SNPs in coding sequences can cause different

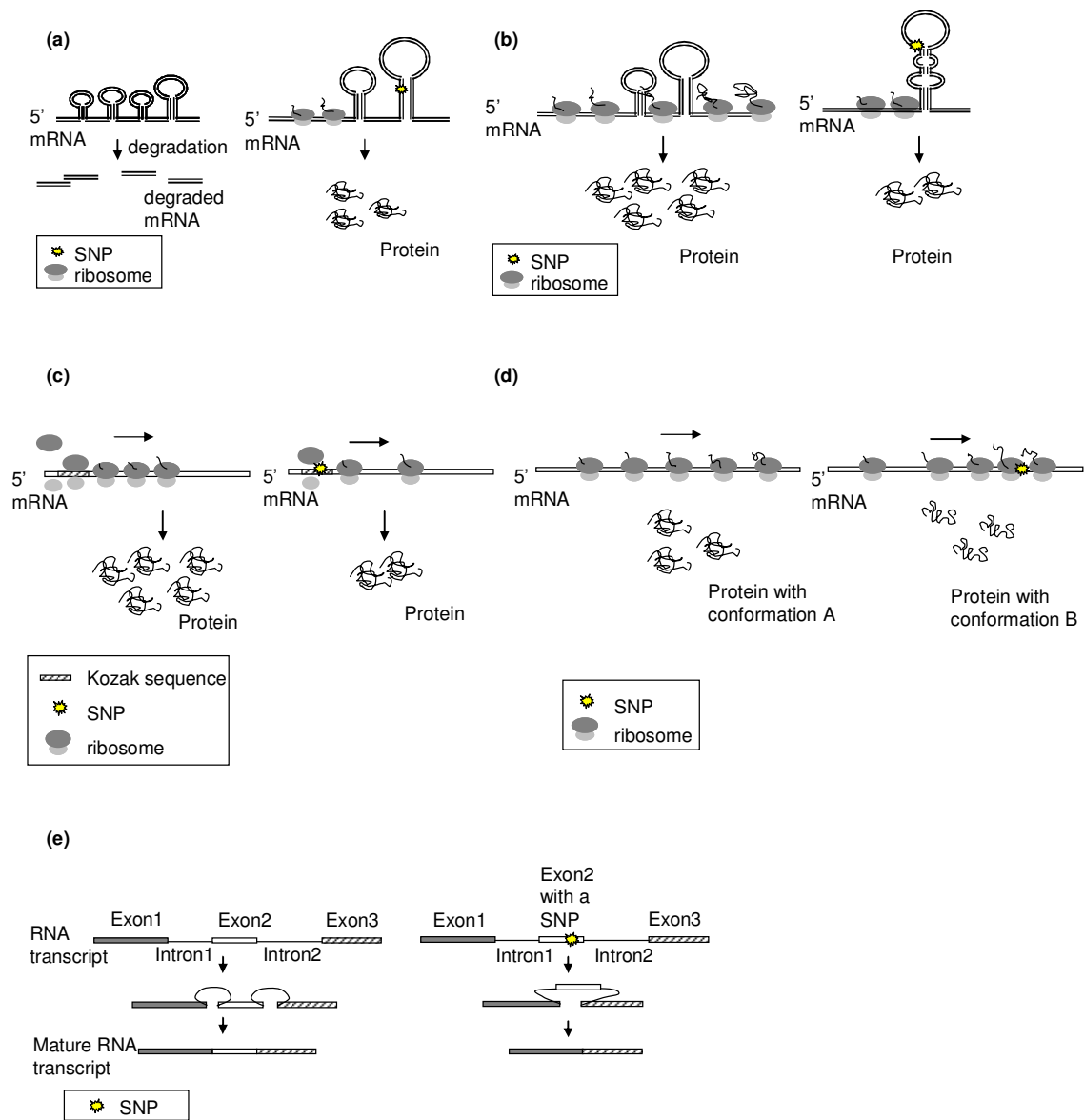


Figure 8.4. Several examples of the effects of SNPs on mRNA and translational levels. (a) A SNP may alter mRNA folding and increase stability, allowing higher level of protein production. (b) A SNP may increase the stability of the mRNA structure, permitting less protein translation. (c) A SNP in Kozak sequence alters the efficiency of translation initiation, resulting in different protein levels. (d) A SNP in the coding sequence changes the kinetics of protein translation, resulting in proteins with different conformations. (e) A SNP in the exon region induces an exon skipping, resulting shorter products.

levels of enzyme activities. Thiopurine S-methyltransferase (TPMT) catalyses the S-methylation of thiopurine drugs. Several human TPMT variant alleles that alter the encoded amino acid sequence of the enzyme generate less stable proteins [42].

Therefore, patients with those alleles have very low TPMT activity and suffer severe, life-threatening drug toxicity when treated with 'standard' doses of thiopurine drugs [43]. A non-synonymous SNP (A428G) in human S-adenosylhomocysteine hydrolase (AdoHcyase) is found in patients with AdoHcyase deficiency [44]. This mutation decreased the unfolding temperature by 7 °C as compared to a wild-type protein. The mutant protein is more sensitive to temperature change and undergoes accelerated aggregation with increasing temperature [45].

Non-synonymous SNPs may alter protein binding affinities and catalytic properties. A SNP is known to alter the regulation of arginine biosynthesis in *E. coli* K12 and B strains [46,47]. In *E. coli* K12, arginine represses the expression of arginine biosynthesis genes and the absence of arginine activates the expression of these genes. In *E. coli* B, these genes are constitutively expressed at low levels regardless of the presence of arginine [48]. These different regulatory patterns result from a SNP at site 70 in the arginine repressor (ArgR) sequence, where the proline of ArgR<sup>K12</sup> is replaced by leucine in ArgR<sup>B</sup> [47]. X-ray crystallography [49,50] indicated that Pro70Leu mutation may alter communications between the DNA-binding domain and the arginine-binding domain by increasing peptide main-chain flexibility. In the absence of arginine, ArgR<sup>B</sup> has a higher binding affinity to the DNA operator than ArgR<sup>K12</sup>. In the presence of arginine, arginine-ArgR<sup>B</sup> has a lower binding affinity to the DNA operator than arginine-ArgR<sup>K12</sup>. As a result, *E. coli* cells carrying ArgR<sup>B</sup> are capable of neither full induction nor complete repression and remain weak constitutive for arginine biosynthesis gene expression, whereas those carrying ArgR<sup>K12</sup> are strongly

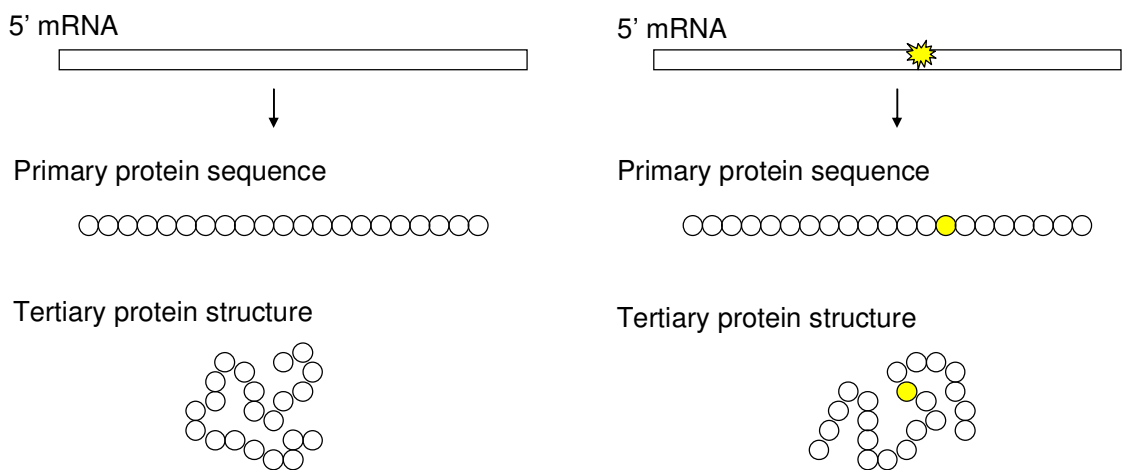


Figure 8.5. An example of the effect of a SNP on the protein and post-transcriptional levels. A non-synonymous SNP alters amino acid sequence, which may change protein conformation. Variations in protein folding may result in different protein stabilities, binding affinities, catalytic properties, and post-translational modifications



regulated. Non-synonymous SNPs in the *fimH* gene can alter the affinity of type 1 fimbriae of uropathogenic *E. coli* for mono-mannose urothelial receptors [51,52]. The stronger adhesion strains conferred increased virulence in the mouse urinary tract. Pregnane X receptor (PXR) is a master transcriptional regulator of many drug/xenobiotic-metabolizing enzymes. Functional analysis of four naturally occurring human PXR variants showed that the Arg98Cys variant altered the DNA binding ability and thus failed to transactivate a reporter containing a PXR responsive element [53]. Melanoma is the most lethal of all skin cancers. Gartside *et al.* reported that 10% of melanoma tumors and cell lines harbor mutations in the fibroblast growth factor receptor 2 gene [54]. With crystal structure mapping, *in vitro* and *in vivo* studies, these mutations were shown to damage protein functions by altering ligand binding affinity, impairing receptor dimerization, destabilizing of the extracellular domains, and reducing kinase activity.

Non-synonymous SNPs can affect post-translational modifications. Because protein phosphorylation is one of the key elements in signal transduction, an altered phosphorylation pattern can cause different responses to the environment. The human ERG1 channel polymorphism is associated with cardiac arrhythmias [55]. Gentile *et al.* reported that a SNP leading to a Lys897Thr substitution creates a phosphorylation site in ERG1, which in turn inhibits channel activity for the downstream signal transduction [56]. Trypsinogen is the most abundant digestive proenzyme produced by the pancreas in human [57]. Ronai *et al.* identified a naturally occurred SNP causing Asp153His substitution in the PRSS2, an anionic trypsinogen [58]. Because Asp153 is the main determinant of tyrosine sulfation in anionic trypsinogen, Asp153His substitution would result in a complete loss of trypsinogen sulfation in PRSS2. Although tyrosine sulfation had no significant effect on the activation of anionic

trypsinogen, the sulfation may play a role in stimulating autoactivation of a cationic trypsinogen, PRSS1.

### ***8.6 Applications to the biotechnology industry***

Although not yet widely applied in the context of biochemical engineering and the biotechnology industry, we believe that knowledge of SNPs may be applied in three ways. First, SNPs are genetic markers for a desired phenotype. For example, most conventional trait markers and molecular markers for animal or crop breeding are based on SNPs. Second, SNPs identified in adaptive evolution help to characterize strains and understand biological pathways. Finally, the molecular effects of SNPs can be used to improve biomolecule production. The following sections will provide examples for each use from the literature that demonstrate the power of SNP knowledge.

#### **8.6.1 Breeding**

SNPs have been extensively used in animal breeding. Because SNPs may alter the expression of the genes participating in inflammation, SNPs are used to define an animal's risk of developing chronic infection. A SNP (C1185T) in IL-10Ralpha is significantly associated with somatic cell score, a trait highly correlated with the incidence of mastitis [59]. In the same study, the haplotype in IL-10Ralpha A-A-T showed higher somatic cell score compared to the most common haplotype. As a result, these markers could serve as risk factors for dairy herd breeding. SNPs can be used as selective markers to improve product quality. Kim *et al.* identified a C-to-T SNP in the 5' UTR of the myogenin gene to associate with muscle fiber and lean meat production in pig breeding [60]. Schennink *et al.* genotyped several candidate genes, *FASN*, *OLR1*, *PPARGC1A*, *PRL* and *STAT5A*, in cows and found that an A-to-G

substitution in *FASN* and a C-to A substitution in *OLRI* have a significant effect on milk-fat percentage [61].

Recently, SNP markers have gained importance in crop breeding. A SNP causing a Lys71Asn substitution in rice alternative oxidase gene was found to contribute to different low temperature tolerance [62]. To obtain effective and durable fungal resistance wheat cultivars, Lagudah *et al.* analyzed variations in the *Lr34* gene [63]. This study identified three polymorphisms, a SNP in intron 4, a 3bp deletion in exon 11 and a SNP in exon 12, to differentiate pathogen susceptible and resistant wheat cultivars. Therefore, these variations can be used as markers to predict pathogen resistance trait in wheat breeding. Wang *et al.* suggested that gene variants can help to determine the shelf life in apples [64]. A SNP in the coding region of the apple *ACS3a* gene causes a Gly289Val substitution in the active site, rendering the enzyme, 1-aminocyclopropane-1-carboxylic acid synthase (ACS), non-functional. This enzyme plays a key role for ethylene production, driving the ripening process. Apple cultivars such as Kitaro and Koukou containing this SNP show much lower ACS activity and maintain fruit firmness for a longer period of time. We believe that the application of SNPs to identify strains of interest, whether animals, rice, apples, or Chinese hamster ovary or *E. coli* cells, represents an important step forward.

### **8.6.2 Strain evolution**

SNPs can provide insight into adaptation processes to the environment. Particularly, several studies have used SNPs to characterize epidemic pathogens. Zhang *et al.* analyzed 1199 chromosomal genes and 92,721 bp of the large virulence plasmid (pO157) of eleven outbreak-associated *E. coli* O157 strains [65]. A total of 906 SNPs in 523 chromosomal genes was identified. The systematic analysis of SNPs is useful

for outbreak investigations, because the result can provide insights to an epidemiologic assessment of associations between bacterial genotypes and disease [66]. *Pseudomonas aeruginosa* causes entailing acute pneumonia and sepsis accompanying with various mortality rates in human [67]. Smith *et al.* performed whole genome sequencing for early and late *P. aeruginosa* isolates [68]. The study revealed several SNPs in the late isolate that were advantageous to for living in hosts. Interestingly, many genes involving in virulence lost their functions in the late isolate, indicating that these gene products may become a burden once the chronic infection has established. The genetic difference between early infectious strains and late adaptive strains may offer new therapeutic opportunities. Influenza viruses can develop antiviral drug resistance, a major challenge for public health epidemiology. Sheu *et al.* identified SNP markers to monitor influenza resistance to two antiviral drugs, oseltamivir and zanamivir [69]. By using SNP markers, they also detected a rise in oseltamivir resistance among A (H1N1) viruses isolated from untreated patients. Because oseltamivir is the most frequently prescribed antiviral agent in the United States for the control of seasonal influenza infections, monitoring the drug resistance by SNPs provides an efficient surveillance system for the public health care agency.

Strain evolution studies based on SNPs are not limited to pathogens. Several non-synonymous SNPs in the MalT activator protein are associated with glucose-limited adaptation in *E. coli* [70]. MalT is a central protein in maltose regulon [71]. The maltose regulon can control the expression of the LamB protein, which, in turn, controls outer membrane permeability for nutrient uptake [72]. In another study of *E. coli* glucose-limited chemostats, a single T-to-A substitution upstream of the gene encoding acetyl CoA synthetase was believed responsible for the semiconstitutive overexpression of this synthetase [73]. The higher level of acetyl CoA synthetase may

contribute to the crossfeeding phenotype, in which one type of cell secretes acetate that another cell can use as a resource in a glucose-limited condition. Herring *et al.* studied the genetic basis of the adaptation to glycerol minimal medium for five *E. coli* populations [74]. The study confirmed 17 SNPs in the five *E. coli* clones. These SNPs lead to several key genes for glycerol adaptation in *E. coli*: all clones had mutations in the gene for glycerol kinase (*glpK*), which catalyzes the first step in glycerol catabolism; mutations in genes encoding the two major subunits of RNA polymerase (*rpoB* and *rpoC*) contributed to 48%–65% of the total change in growth rate for glycerol adaptation. These results provide insight for glycerol metabolism in *E. coli*. Interestingly, there are SNP hotspots in core genes of *E. coli* under short term positive selection [75]. This observation suggests that mutations are likely to occur in specific position in genes to modify the function of encoded proteins in a specific, fine-tuned manner. Taken together, strain evolution studies based on SNPs analysis offer a systematic methodology to identify candidate genes for future manipulations.

### **8.6.3 Biomolecule production**

Because SNPs can affect gene regulation and protein function, knowledge of functional SNPs can be applied to genetic engineering to create phenotypes of interest. To enhance an extracellular secretion system in *E. coli*, a hypersecreter strain (B41) was created by chemical mutagenesis [76]. B41 strain was found to secrete four-fold more hemolysin (HlyA) protein relative to the parent strain via the Type I secretion pathway. The genomes of the parent and the B41 strains were later sequenced by a Illumina Genome Analyzer [77]. A G-to-T substitution in *ycdC* gene was found in B41 strain but not in the parent strain. This SNP caused a premature termination of the encoded protein, RutR, suggesting a role of RutR on Type I secretion pathway in *E. coli*. This study demonstrates the usage of SNP analysis to characterize a recombinant

host, allowing better phenotype predictions in the future.

The detoxified beta-toxoid is a putative vaccine component, which can be produced and secreted by *Bacillus subtilis*. However, the secretion yield for beta-toxoid is low because of rapid degradation. Nijland *et al.* identified a single amino acid substitution in the beta-toxoid sequence to alter protein folding, which slows down the degradation and increases secretion yield [78]. This result demonstrates that in addition to host adaptation, the intrinsic properties of a heterologous protein can affect its own productions.

The Sindbis viral expression system provides a rapid production of recombinant protein in mammalian cells [79]. This expression, however, is typically limited to transient production because of the cytotoxicity of the virus. A SNP (C3855T) leading to a Phe726Ser substitution in a non-structural viral protein, nsP2, was found to reduce viral cytotoxicity [80-82]. A C3856T variant causing Phe726Leu mutation in the same protein also results in noncytopathic infection [83]. As a result, Phe726Leu and Phe726Ser mutations were engineered into a Sindbis virus replicon to enable a sustained expression of recombinant proteins [81-84].

Recombinant adeno-associated virus 2 (AAV2) vectors are used as gene delivery vehicles in several Phase I/II clinical trials, but relatively large vector doses are needed to achieve therapeutic benefits [85]. To reduce dose while achieving similar clinical benefit, a point mutation causing Tyr-to-Phe substitution was introduced into the AAV2 vector [86]. This mutation helps the AAV2 prevent capsid ubiquitination and improve intracellular trafficking to the nucleus. The tyrosine mutant vector achieves therapeutic levels of human Factor IX at an about 10-fold reduced vector dose. This

study provides a strategy using single nucleotide substitution to improve human gene therapy.

Polyhydroxyalkanoate synthase (PhaC) has been introduced into *E. coli* to produce biopolymer, poly(3-hydroxybutyrate) (PHB). It was found that PhaC with a Gly4Asp substitution resulted in higher level of PhaC and PHB productions [87]. This mutation was further applied to enhance PHB production in *Corynebacterium glutamicum* [88].

### ***8.7 Challenges and the future***

Currently, genetic engineering in biotechnology is mostly based on candidate gene mutagenesis and screening. As mentioned above, candidate gene studies rely on prior knowledge of biological pathways and are less likely to discover novel gene targets. On the other hand, genome-wide association studies have identified SNPs in novel gene targets associated with human disease and quantitative traits. In addition, genotyping technologies have also been applied to characterize epidemic microorganisms. Recent attempts to identify SNPs for differentiating regional *Mycobacterium ulcerans* strains demonstrate the strength of a genome-wide SNP analysis [89]. These biomedical studies will provide insight into SNPs analysis for strain developments in biotechnology applications.

The lack of SNP databases for commonly used recombinant hosts such as *E. coli*, *B. subtilis* and Chinese hamster ovary cells (CHO) makes genome-wide scale SNP analysis difficult at this time. Next generation sequencing technology allows SNP detections and novel SNP discoveries. For example, Monsanto has applied next generation sequencing platform to identify SNPs and to characterize genetic variation in maize lines. A United States Department of Agriculture study used an Illumina

Genome analyzer to genotype 66 cattle, identifying and validating approximately 23,000 new bovine SNPs [90]. This research led to a bovine genotyping array now commercialized by Illumina. Importantly, for recombinant hosts generated by mutagenesis, SNPs may occur at any position in the genome. Thus a chip-based detection for known genetic variations is less applicable and the next generation sequencing may be a stronger tool in detecting random mutations.

Another challenge for CHO cell SNPs analysis is the lack of a reference genome. Reduced complexity approaches such as EST sequencing can provide adequate sequence depth for SNP discovery without sampling the complete genome. To analyze the uncharacterized *Eucalyptus grandis* genome, Novaes *et al.* used Roche 454 technology to sequence and assemble 148 Mbp EST sequences [91]. By aligning sequencing reads from multiple genotypes, 23,742 SNPs were predicted, 83% of which were validated. In addition, it is also possible to combine long-read and short-read next generation sequencing to identify SNPs for species with no available reference genomes. Importantly, Wlaschin *et al.* has established EST libraries for the CHO cells [92], and the same group later reported a scaffold for the CHO genome [93]. Recently, bacterial artificial chromosome libraries were created to further characterize the CHO genome [94]. It is important to note that, however, EST based approaches cannot address the issue of SNPs in noncoding regions of the genome. As the knowledge of the CHO genome grows, genome wide scale SNP analysis will become more feasible.

SNP detection is an unexplored tool for improving and accelerating high-producer cell line generation and selection. Although not yet applied to biotechnology platform organisms, the availability of next generation sequencing technology provides a



significant opportunity to allow this community to embrace a new approach to cell line characterization and process platform development. While large-scale analysis is initially based on individual SNPs, the focus will soon shift to haplotype-specific SNPs for more efficient association studies as it is done in the human genome analysis. Haplotype-based SNPs identified in recombinant hosts will open up an efficient management of genetic diversity in the strain development on a whole genome level.

## REFERENCES

1. Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J., et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077-1082.
2. The International HapMap Consortium. (2003) The international HapMap Project, *Nature*, **426**, 789-796.
3. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299-1320.
4. The International HapMap Consortium (2007) A second-generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-61.
5. Rafalski,A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.*, **5**, 94-100.
6. Ganai,M.W., Altmann,T. and Roder,M.S. (2009) SNP identification in crop plants. *Curr. Opin. Plant Biol.*, **12**, 211-217.
7. Kim,S. and Misra,A. (2007) SNP genotyping: Technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, **9**, 289-320.
8. Imelfort,M., Duran,C., Batley,J. and Edwards,D. (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant. Biotechnol. J.*, **7**, 312-317.
9. Smith,E., Shi,L., Drummond,P., Rodriguez,L., Hamilton,R., Powell,E., Nahashon,S., Ramlal,S., Smith,G. and Foster,J. (2000) Development and characterization of expressed sequence tags for the turkey (*Meleagris gallopavo*) genome and comparative sequence analysis with other birds. *Anim. Genet.*, **31**, 62-67.
10. He,C., Chen,L., Simmons,M., Li,P., Kim,S. and Liu,Z.J. (2003) Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Anim. Genet.*, **34**, 445-448.
11. Somers,D.J., Kirkpatrick,R., Moniwa,M. and Walsh,A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome*, **46**, 431-437.
12. Mardis,E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133-141.

13. Tabor,H.K., Risch,N.J. and Myers,R.M. (2002) Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nat. Rev. Genet.*, **3**, 391-397.
14. Pearson,T.A. and Manolio,T.A. (2008) How to interpret a genome-wide association study. *JAMA*, **299**, 1335-1344
15. Nickerson,D.A., Taylor,S.L., Weiss,K.M., Clark,A.G., Hutchinson,R.G., Stengard,J., Salomaa,V., Vartiainen,E., Boerwinkle,E. and Sing,C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.*, **19**, 233-240.
16. Chamary,J.V., Parmley,J.L. and Hurst,L.D. (2006) Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98-108.
17. Sauna,Z.E., Kimchi-Sarfaty,C., Ambudkar,S.V. and Gottesman,M.M. (2007) Silent polymorphisms speak: How they affect pharmacogenomics and the treatment of cancer. *Cancer Res.*, **67**, 9609-9612.
18. Gupta,P. and Lee,K.H. (2008) Silent mutations result in HlyA hypersecretion by reducing intracellular HlyA protein aggregates. *Biotechnol. Bioeng.*, **101**, 967-974.
19. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255-258.
20. Mooney,S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform*, **6**, 44-56.
21. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.*, **12**, 2249-2254.
22. Giacomelli,F., Marciano,R., Pistorio,A., Catarsi,P., Canini,S., Karsenty,G. and Ravazzolo,R. (2004) Polymorphisms in the osteopontin promoter affect its transcriptional activity. *Physiol. Genomics*, **20**, 87-96.
23. Gudmundsson,J., Sulem,P., Gudbjartsson,D.F., Jonasson,J.G., Sigurdsson,A., Bergthorsson,J.T., He,H., Blondal,T., Geller,F., Jakobsdottir,M., et al. (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in european populations. *Nat. Genet.*, **41**, 460-464.
24. Landa,I., Ruiz-Llorete,S., Montero-Conde,C., Inglada-Perez,L., Schiavi,F., Leskela,S., Pita,G., Milne,R., Maravall,J., Ramos,I., et al. (2009) The variant rs1867277 in *FOXE1* gene confers thyroid cancer susceptibility through the recruitment of USF1/USF2 transcription factors. *PLoS Genet.*, **5**, e1000637.

25. Egli,R.J., Southam,L., Wilkins,J.M., Lorenzen,I., Pombo-Suarez,M., Gonzalez,A., Carr,A., Chapman,K. and Loughlin,J. (2009) Functional analysis of the osteoarthritis susceptibility-associated GDF5 regulatory polymorphism. *Arthritis Rheum.*, **60**, 2055-2064.
26. Duan,J., Wainwright,M.S., Comeron,J.M., Saitou,N., Sanders,A.R., Gelernter,J. and Gejman,P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, **12**, 205-216.
27. Capon,F., Allen,M.H., Ameen,M., Burden,A.D., Tillman,D., Barker,J.N. and Trembath,R.C. (2004) A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.*, **13**, 2361-2368.
28. Chen,G.L. and Miller,G.M. (2008) Rhesus monkey tryptophan hydroxylase-2 coding region haplotypes affect mRNA stability. *Neuroscience*, **155**, 485-491.
29. Nackley,A.G., Shabalina,S.A., Tchivileva,I.E., Satterfield,K., Korchynskyi,O., Makarov,S.S., Maixner,W. and Diatchenko,L. (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, **314**, 1930-1933.
30. Diatchenko,L., Slade,G.D., Nackley,A.G., Bhalang,K., Sigurdsson,A., Belfer,I., Goldman,D., Xu,K., Shabalina,S.A., Shagin,D., et al. (2005) Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Hum. Mol. Genet.*, **14**, 135-143.
31. Jacobson,E.M., Concepcion,E., Oashi,T. and Tomer,Y. (2005) A graves' disease-associated kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: A case for translational pathophysiology. *Endocrinology*, **146**, 2684-2691.
32. Komar,A.A., Lesnik,T. and Reiss,C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation. *FEBS Lett.*, **462**, 387-391.
33. Kimchi-Sarfaty,C., Oh,J.M., Kim,I.W., Sauna,Z.E., Calcagno,A.M., Ambudkar,S.V. and Gottesman,M.M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, **315**, 525-528.
34. Wachtel,C. and Manley,J.L. (2009) Splicing of mRNA precursors: The role of RNAs and proteins in catalysis. *Mol. Biosyst*, **5**, 311-316.
35. Pagani,F., Stuani,C., Tzetis,M., Kanavakis,E., Efthymiadou,A., Doudounakis,S., Casals,T. and Baralle,F.E. (2003) New type of disease causing mutations: The

example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum. Mol. Genet.*, **12**, 1111-1120.

36. Nicolas,F.E., Lopez-Gomollon,S., Lopez-Martinez,A.F. and Dalmay,T. (2009) RNA silencing: Recent developments on miRNAs. *Recent. Pat. DNA Gene Seq*, **3**, 77-87.
37. Gottwein,E., Cai,X. and Cullen,B.R. (2006) A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects drosha processing. *J. Virol.*, **80**, 5321-5326.
38. Sun,G., Yan,J., Noltner,K., Feng,J., Li,H., Sarkis,D.A., Sommer,S.S. and Rossi,J.J. (2009) SNPs in human miRNA genes affect biogenesis and function. *RNA*, **15**, 1640-1651.
39. Wang,Z. and Moulton,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263-270.
40. Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61-80.
41. Teng,S., Michonova-Alexova,E. and Alexov,E. (2008) Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Curr. Pharm. Biotechnol.*, **9**, 123-133.
42. Salavaggione,O.E., Wang,L., Wiepert,M., Yee,V.C. and Weinshilboum,R.M. (2005) Thiopurine S-methyltransferase pharmacogenetics: Variant allele functional and comparative genomics. *Pharmacogenet Genomics*, **15**, 801-815.
43. Lennard,L., Van Loon,J.A. and Weinshilboum,R.M. (1989) Pharmacogenetics of acute azathioprine toxicity: Relationship to thiopurine methyltransferase genetic polymorphism. *Clin. Pharmacol. Ther.*, **46**, 149-154.
44. Baric,I., Cuk,M., Fumic,K., Vugrek,O., Allen,R.H., Glenn,B., Maradin,M., Pazanin,L., Pogribny,I., Rados,M., et al. (2005) S-adenosylhomocysteine hydrolase deficiency: A second patient, the younger brother of the index patient, and outcomes during therapy. *J. Inherit. Metab. Dis.*, **28**, 885-902.
45. Beluzic,R., Cuk,M., Pavkov,T., Fumic,K., Baric,I., Mudd,S.H., Jurak,I. and Vugrek,O. (2006) A single mutation at Tyr143 of human S-adenosylhomocysteine hydrolase renders the enzyme thermosensitive and affects the oxidation state of bound cofactor nicotinamide-adenine dinucleotide. *Biochem. J.*, **400**, 245-253.
46. Lim,D.B., Oppenheim,J.D., Eckhardt,T. and Maas,W.K. (1987) Nucleotide sequence of the argR gene of *Escherichia coli* K-12 and isolation of its product, the arginine repressor. *Proc. Natl. Acad. Sci. U. S. A.*, **84**, 6697-6701.

47. Lim,D., Oppenheim,J.D., Eckhardt,T. and Maas,W.K. (1988) The unitary hypothesis for the repression mechanism of arginine biosynthesis in *E. coli* B and *E. coli* K12—revisited after 18 years. In Bissel,M., Deho,G., Sironi,G. and Torriani,A. (eds.), *Gene Expression and Regulation: the Legacy of Luigi Gorinini*. Excerpta Medica, New York, pp. 55-63.
48. Jacoby,G.A. and Gorini,L. (1967) Genetics of control of the arginine pathway in *Escherichia coli* B and K. *J. Mol. Biol.*, **24**, 41-50.
49. Van Duyne,G.D., Ghosh,G., Maas,W.K. and Sigler,P.B. (1996) Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*. *J. Mol. Biol.*, **256**, 377-391.
50. Sunnerhagen,M., Nilges,M., Otting,G. and Carey,J. (1997) Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA. *Nat. Struct. Biol.*, **4**, 819-826.
51. Sokurenko,E.V., Chesnokova,V., Dykhuizen,D.E., Ofek,I., Wu,X.R., Krogfelt,K.A., Struve,C., Schembri,M.A. and Hasty,D.L. (1998) Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 8922-8926.
52. Pouttu,R., Puustinen,T., Virkola,R., Hacker,J., Klemm,P. and Korhonen,T.K. (1999) Amino acid residue ala-62 in the FimH fimbrial adhesin is critical for the adhesiveness of meningitis-associated *Escherichia coli* to collagens. *Mol. Microbiol.*, **31**, 1747-1757.
53. Koyano,S., Kurose,K., Saito,Y., Ozawa,S., Hasegawa,R., Komamura,K., Ueno,K., Kamakura,S., Kitakaze,M., Nakajima,T., et al. (2004) Functional characterization of four naturally occurring variants of human pregnane X receptor (PXR): One variant causes dramatic loss of both DNA binding activity and the transactivation of the CYP3A4 promoter/enhancer region. *Drug Metab. Dispos.*, **32**, 149-154.
54. Gartside,M.G., Chen,H., Ibrahimi,O.A., Byron,S.A., Curtis,A.V., Wellens,C.L., Bengston,A., Yudt,L.M., Eliseenkova,A.V., Ma,J., et al. (2009) Loss-of-function fibroblast growth factor receptor-2 mutations in melanoma. *Mol. Cancer. Res.*, **7**, 41-54.
55. Curran,M.E., Splawski,I., Timothy,K.W., Vincent,G.M., Green,E.D. and Keating,M.T. (1995) A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. *Cell*, **80**, 795-803.
56. Gentile,S., Martin,N., Scappini,E., Williams,J., Erxleben,C. and Armstrong,D.L. (2008) The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 14704-14708.

57. Chen,J.M. and Férec,C. (2003) Trypsinogen genes: evolution. In Cooper,D.N. (ed), *Nature encyclopedia of the human genome*. Macmillan, London, pp. 645–650.
58. Ronai,Z., Witt,H., Rickards,O., Destro-Bisol,G., Bradbury,A.R. and Sahin-Toth,M. (2009) A common african polymorphism abolishes tyrosine sulfation of human anionic trypsinogen (PRSS2). *Biochem. J.*, **418**, 155-161.
59. Verschoor,C.P., Pant,S.D., Schenkel,F.S., Sharma,B.S. and Karrow,N.A. (2009) SNPs in the bovine IL-10 receptor are associated with somatic cell score in canadian dairy bulls. *Mamm. Genome*, **20**, 447-454.
60. Kim,J.M., Choi,B.D., Kim,B.C., Park,S.S. and Hong,K.C. (2009) Associations of the variation in the porcine myogenin gene with muscle fibre characteristics, lean meat production and meat quality traits. *J. Anim. Breed. Genet.*, **126**, 134-141.
61. Schennink,A., Bovenhuis,H., Leon-Kloosterziel,K.M., van Arendonk,J.A. and Visker,M.H. (2009) Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. *Anim. Genet.*, doi:10.1111/j.1365-2052.2009.01940.
62. Abe,F., Saito,K., Miura,K. and Toriyama,K. (2002) A single nucleotide polymorphism in the alternative oxidase gene among rice varieties differing in low temperature tolerance. *FEBS Lett.*, **527**, 181-185.
63. Lagudah,E.S., Krattinger,S.G., Herrera-Foessel,S., Singh,R.P., Huerta-Espino,J., Spielmeyer,W., Brown-Guedira,G., Selter,L.L. and Keller,B. (2009) Gene-specific markers for the wheat gene Lr34/Yr18/Pm38 which confers resistance to multiple fungal pathogens. *Theor. Appl. Genet.*, **119**, 889-898.
64. Wang,A., Yamakake,J., Kudo,H., Wakasa,Y., Hatsuyama,Y., Igarashi,M., Kasai,A., Li,T. and Harada,T. (2009) Null mutation of the *MdACS3* gene, coding for a ripening-specific 1-aminocyclopropane-1-carboxylate synthase, leads to long shelf life in apple fruit. *Plant Physiol.*, **151**, 391-399.
65. Zhang,W., Qi,W., Albert,T.J., Motiwala,A.S., Alland,D., Hyytia-Trees,E.K., Ribot,E.M., Fields,P.I., Whittam,T.S. and Swaminathan,B. (2006) Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. *Genome Res.*, **16**, 757-767.
66. Manning,S.D., Motiwala,A.S., Springman,A.C., Qi,W., Lacher,D.W., Ouellette,L.M., Mladonicky,J.M., Somsel,P., Rudrik,J.T., Dietrich,S.E., et al. (2008) Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 4868-4873.

67. El Solh, A.A., Akinnusi, M.E., Wiener-Kronish, J.P., Lynch, S.V., Pineda, L.A. and Szarpa, K. (2008) Persistent infection with *Pseudomonas aeruginosa* in ventilator-associated pneumonia. *Am. J. Respir. Crit. Care Med.*, **178**, 513-519.
68. Smith, E.E., Buckley, D.G., Wu, Z., Saenphimmachak, C., Hoffman, L.R., D'Argenio, D.A., Miller, S.I., Ramsey, B.W., Speert, D.P., Moskowitz, S.M., et al. (2006) Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 8487-8492.
69. Sheu, T.G., Deyde, V.M., Okomo-Adhiambo, M., Garten, R.J., Xu, X., Bright, R.A., Butler, E.N., Wallis, T.R., Klimov, A.I. and Gubareva, L.V. (2008) Surveillance for neuraminidase inhibitor resistance among human influenza A and B viruses circulating worldwide from 2004 to 2008. *Antimicrob. Agents Chemother.*, **52**, 3284-3292.
70. Notley-McRobb, L. and Ferenci, T. (1999) The generation of multiple co-existing mal-regulatory mutations through polygenic evolution in glucose-limited populations of *Escherichia coli*. *Environ. Microbiol.*, **1**, 45-52.
71. Schwartz, M. (1987) The maltose regulon. In Neidhardt, F.C., Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M. and Umberger, H.E. (eds), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. American Society for Microbiology Press, Washington DC, pp. 1482-1502.
72. Death, A. and Ferenci, T. (1993) The importance of the binding-protein-dependent *mgI* system to the transport of glucose in *Escherichia coli* growing on low sugar concentrations. *Res. Microbiol.*, **144**, 529-537.
73. Treves, D.S., Manning, S. and Adams, J. (1998) Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Mol. Biol. Evol.*, **15**, 789-797.
74. Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R., Albert, T.J., Blattner, F.R., van den Boom, D., Cantor, C.R., et al. (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.*, **38**, 1406-1412.
75. Chattopadhyay, S., Weissman, S.J., Minin, V.N., Russo, T.A., Dykhuizen, D.E. and Sokurenko, E.V. (2009) High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 12412-12417.
76. Lee, P.S. and Lee, K.H. (2005) Engineering HlyA hypersecretion in *Escherichia coli* based on proteomic and microarray analyses. *Biotechnol. Bioeng.*, **89**, 195-205.
77. Gupta, P., Swanberg, J.C. and Lee, K.H. submitted.



78. Nijland,R., Heerlien,R., Hamoen,L.W. and Kuipers,O.P. (2007) Changing a single amino acid in clostridium perfringens beta-toxin affects the efficiency of heterologous secretion by bacillus subtilis. *Appl. Environ. Microbiol.*, **73**, 1586-1593.
79. Mastrangelo,A.J., Hardwick,J.M., Bex,F. and Betenbaugh,M.J. (2000) Part I. bcl-2 and bcl-x(L) limit apoptosis upon infection with alphavirus vectors. *Biotechnol. Bioeng.*, **67**, 544-554.
80. Dryga,S.A., Dryga,O.A. and Schlesinger,S. (1997) Identification of mutations in a sindbis virus variant able to establish persistent infection in BHK cells: The importance of a mutation in the nsP2 gene. *Virology*, **228**, 74-83.
81. Nivitchanyong,T., Tsai,Y.C., Betenbaugh,M.J. and Oyler,G.A. (2009) An improved *in vitro* and *in vivo* sindbis virus expression system through host and virus engineering. *Virus Res.*, **141**, 1-12.
82. Kim,J., Dittgen,T., Nimmerjahn,A., Waters,J., Pawlak,V., Helmchen,F., Schlesinger,S., Seeburg,P.H. and Osten,P. (2004) Sindbis vector SINrep(nsP2S726): A tool for rapid heterologous expression with attenuated cytotoxicity in neurons. *J. Neurosci. Methods*, **133**, 81-90.
83. Frolov,I., Agapov,E., Hoffman,T.A.,Jr, Pragai,B.M., Lipka,M., Schlesinger,S. and Rice,C.M. (1999) Selection of RNA replicons capable of persistent noncytopathic replication in mammalian cells. *J. Virol.*, **73**, 3854-3865.
84. Agapov,E.V., Frolov,I., Lindenbach,B.D., Pragai,B.M., Schlesinger,S. and Rice,C.M. (1998) Noncytopathic sindbis virus RNA vectors for heterologous gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 12989-12994.
85. Snyder,R.O. and Francis,J. (2005) Adeno-associated viral vectors for clinical gene transfer studies. *Curr. Gene Ther.*, **5**, 311-321.
86. Zhong,L., Li,B., Mah,C.S., Govindasamy,L., Agbandje-McKenna,M., Cooper,M., Herzog,R.W., Zolotukhin,I., Warrington,K.H.,Jr, Weigel-Van Aken,K.A., et al. (2008) Next generation of adeno-associated virus 2 vectors: Point mutations in tyrosines lead to high-efficiency transduction at lower doses. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 7827-7832.
87. Wong,H.H. and Lee,S.Y. (1998) Poly-(3-hydroxybutyrate) production from whey by high-density cultivation of recombinant *Escherichia coli*. *Appl. Microbiol. Biotechnol.*, **50**, 30-33.
88. Jo,S.J., Matsumoto,K., Leong,C.R., Ooi,T. and Taguchi,S. (2007) Improvement of poly(3-hydroxybutyrate) [P(3HB)] production in *Corynebacterium glutamicum* by

- codon optimization, point mutation and gene dosage of P(3HB) biosynthetic genes. *J. Biosci. Bioeng.*, **104**, 457-463.
89. Kaser,M., Hauser,J. and Pluschke,G. (2009) Single nucleotide polymorphisms on the road to strain differentiation in *Mycobacterium ulcerans*. *J. Clin. Microbiol.*,doi:10.1128/JCM.00761-09.
  90. Van Tassell,C.P., Smith,T.P., Matukumalli,L.K., Taylor,J.F., Schnabel,R.D., Lawley,C.T., Haudenschild,C.D., Moore,S.S., Warren,W.C. and Sonstegard,T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods*, **5**, 247-252.
  91. Novaes,E., Drost,D.R., Farmerie,W.G., Pappas,G.J.,Jr, Grattapaglia,D., Sederoff,R.R. and Kirst,M. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.
  92. Wlaschin,K.F., Nissom,P.M., Gatti Mde,L., Ong,P.F., Arleen,S., Tan,K.S., Rink,A., Cham,B., Wong,K., Yap,M., et al. (2005) EST sequencing for gene discovery in chinese hamster ovary cells. *Biotechnol. Bioeng.*, **91**, 592-606.
  93. Wlaschin,K.F. and Hu,W.S. (2007) A scaffold for the chinese hamster genome. *Biotechnol. Bioeng.*, **98**, 429-439.
  94. Omasa,T., Cao,Y., Park,J.Y., Takagi,Y., Kimura,S., Yano,H., Honda,K., Asakawa,S., Shimizu,N. and Ohtake,H. (2009) Bacterial artificial chromosome library for genome-wide analysis of chinese hamster ovary cells. *Biotechnol. Bioeng.*,**104**, 986-994

## CHAPTER 9

### CONCLUSION AND FUTURE DIRECTIONS

#### ***9.1 Summary***

This research dissertation focuses on developing computational and experimental tools to understand the mechanism of +1 PRF and -1 PRF. The objective is to find a systematic way to analyze a complex, but critically important, biological event.

Three computational programs were developed in this dissertation project. First, a kinetic model was developed for +1 PRF. The model demonstrated that stimulatory signals leading to the release of deacylated tRNA in the E-site, a slippery site allowing the P-site tRNA to re-pair with the new reading frame, and an A-site codon with a low aminoacyl-tRNA concentration synergistically promoted efficient +1 frameshifting. Second, motivated by the +1 PRF model prediction, a bioinformatic program, FSscan, was constructed to search +1 frameshift hot spots in the *Escherichia coli* genome. FSscan calculated scores for a 16-nucleotide window along a gene sequence according to different effects of the stimulatory signals, and ribosome E-, P-, and A-site interactions. FSscan predicted *yehP*, *pepP*, *nuoE*, and *cheA* as +1 frameshift candidates in the *E. coli* genome. Finally, a kinetic model was built for -1 PRF. The model yielded two possible -1 frameshift products: those incorporating zero frame A-site tRNAs in the recoding site and products incorporating -1 frame A-site tRNAs in the recoding site. The model calculated not only the change in frameshift efficiency, but also the change in the composition of frameshift products under different conditions. In addition, the model identified high impact parameters, representing steps in the translation elongation, on -1 frameshifting.

For experimental tools, a dual fluorescence reporter system was developed in *E. coli* and yeast *S. cerevisiae*. Different PRF sites were inserted between two fluorescence reporter genes, monomeric DsRed and EGFP. The red and green fluorescence for different strains were directly measured *in vivo* by a microwell plate reader. The system allows an easy comparison of frameshift efficiency for different recoding sites with the normalized fluorescence ratio because the assay requires neither cell lysis nor additional chemical reactions. In addition, frameshift proteins derived from this dual fluorescence reporter can be further purified and analyzed by mass spectrometry. Using the dual fluorescence reporter system in *E. coli* to study RF2 frameshifting, higher +1 frameshift efficiency was observed for a E-site codon with a weaker codon:anticodon interaction. When examining candidate genes by FSscan, sequences from *yehP*, *pepP*, *nuoE*, and *cheA* revealed +1 frameshift efficiency significantly higher than a randomly design sequence. Experimentally targeting steps in the -1 PRF model resulted in different levels of frameshift efficiency, which were consistent with model predictions. The -1 frameshift proteins were further purified and analyzed by mass spectrometry, empirically demonstrating the fraction of the two types of frameshift products.

## ***9.2 Future directions***

### **9.2.1 Dual fluorescence reporter system in mammalian cells for therapeutic screening**

The dual fluorescence reporter systems in *E. coli* and yeast *S. cerevisiae* allow fast screening for the effect of different genetic backgrounds or chemicals on PRF. However, to develop an antiviral therapeutic screening method, a proper host cell line provides more relevant information. For example, the PRF motif in HIV-1 is translated by a human ribosome in nature. Conditions altering bacterial or yeast

ribosomes in translating PRF signals may not affect human ribosomes in doing so. Plant *et al.* compared HIV-1 frameshifting in *E. coli* and human T-cell and found that phylogenetic differences in ribosome structure can affect frameshift efficiency [1]. Recently, a dual fluorescence reporter has been developed in the mammalian system, COS-7 and HEK 293T cells [2], supporting the high potential of the dual fluorescence reporter system for an anti-viral drug screening. In the present dissertation, several combinations of the fluorescent reporter proteins have been constructed. These reporters can be modified and integrated into the genome of the mammalian cells, preferably a human T-cell line (Jurkat cells). The antiviral therapeutic screening can then be done in a 24-, 96- or 384-well microplate format. In addition, the reporter cell lines will allow a library screening using fluorescence-activated cell sorting (FACS). In an *E. coli* system, Dulude *et al.* constructed a RNA binding peptide library and co-expressed with a dual fluorescence reporter in a separate plasmid. The study reported candidate peptides that reduced frameshift efficiency by 50% [3]. Interestingly, Olsthoorn *et al.*, observed that a small 12nt- to 13nt-RNA complementary to the downstream of a slippery site stimulated -1 ribosomal frameshifting from 0.4% to ~15% in an *in vitro* translation system [4]. Oligonucleotides annealed to the downstream of a shift-prone site UCC UGA were also shown to enhance +1 frameshifting *in vitro* and in mammalian culture cells [5]. Because cell growth can be monitored along with the *in vivo* dual fluorescence reporter assay, the system provides a platform to identify compounds targeting only the PRF without damaging the host. Specifically, several potential therapeutic strategies can be addressed further:

- (1) Screening current Food and Drug Administration approved drugs as well as potential chemicals that may affect PRF in the mammalian cells. Which drugs reveal high potential? Once the candidates are identified, it is worthwhile to investigate how these drugs affect PRF.

(2) Establishing a general rule for a peptide which can influence PRF in the mammalian culture. How long should the peptide be? What types of the amino acid should be incorporate? Will the peptide target all PRF signal or is it specific to one type of the signal?

(3) Developing a strategy to design the antisense to affect PRF in the mammalian culture. How long should the antisense be? What should be the distance between the antisense and the slippery site? Is there an optimal time to induce the antisense expression?

The answers to these questions would provide insight into the drug development for viral diseases or other PRF related disorders.

### **9.2.2 Genome-wide scale PRF cassette identification**

The dual fluorescence reporter system allows an easy *in vivo* assay for PRF efficiency. This system may be used for a genome-wide screening for PRF cassettes. A genomic library can be inserted as a linker sequence between the two fluorescent reporters. Using FACS, cells with relatively high green to red fluorescence ratio are potential candidates containing PRF sequences. The major challenge for this study is the control of the reading frame. Any DNA fragment in the library containing one or two more nucleotides can change the reading frame of *egfp* and affect the result. Therefore, a DNA purification method at a single nucleotide resolution is required. However, this method has not been reported in the literature. Alternatively, one can design a set of the reporter to account for the three reading frame (pRGlib0, pRGlib1 and pRGlib2 in Figure 9.1). Construction of a genomic DNA library by endonuclease digestion restricts the DNA fragmentation pattern. To ensure that each fragment will have the opportunity to combine with a downstream *egfp* in different frames, a library should be cloned into all three reporters. In this system, various lengths of the DNA fragment

would result in different fluorescence patterns. For example, if a DNA fragment renders *egfp* in the zero frame, the intensity of the green fluorescence is expected to be high. If a DNA fragment renders *egfp* in the +1 frame, +1 PRF candidates may reveal the green fluorescence higher than a background. These patterns are summarized at the bottom of Figure 9.1. The system assumes that for PRF-dependent *egfp* expression, the green to red fluorescence ratio is lower than an in-frame *egfp*. Therefore, cells with a medium level of green fluorescence are more interesting candidates. On the other hand, this assumption may reject PRF candidates with high efficiency. As a result, selecting cells within a range of the fluorescence pattern to investigate further become a critical issue.

### 9.2.3 Investigating *yehP*, *pepP*, *nuoE*, and *cheA*

In this work, FSscan identified several +1 PRF candidates, *yehP*, *pepP*, *nuoE*, and *cheA* in the *E. coli* genome. These genes were identified because they contained sequences with a potential to interact with anti-SD sequence in the 16S ribosome, a weak E-site interaction, a slippery P-site, and a hungry A-site codon. Whether an mRNA secondary is involved in these PRF is not clear. A saturation mutagenesis study for each candidate sequence may reveal other novel PRF enhancing elements.

+1 PRF occurring in *yehP*, *pepP*, and *cheA* sequences will result in a shorter polypeptides and +1 PRF occurring in the *nuoE* sequence would result in a longer polypeptide compared to non-frameshift proteins. To date, the function of YehP is not known. A *yehP* knockout *E. coli* strain was previously shown to result in a different swarming phenotype [6]. Other candidate genes have a better known function: *nuoE* codes for NADH:ubiquinone oxidoreductase, chain E, *pepP* codes for proline aminopeptidase P II, and *cheA* codes for a histidine protein kinase sensor of

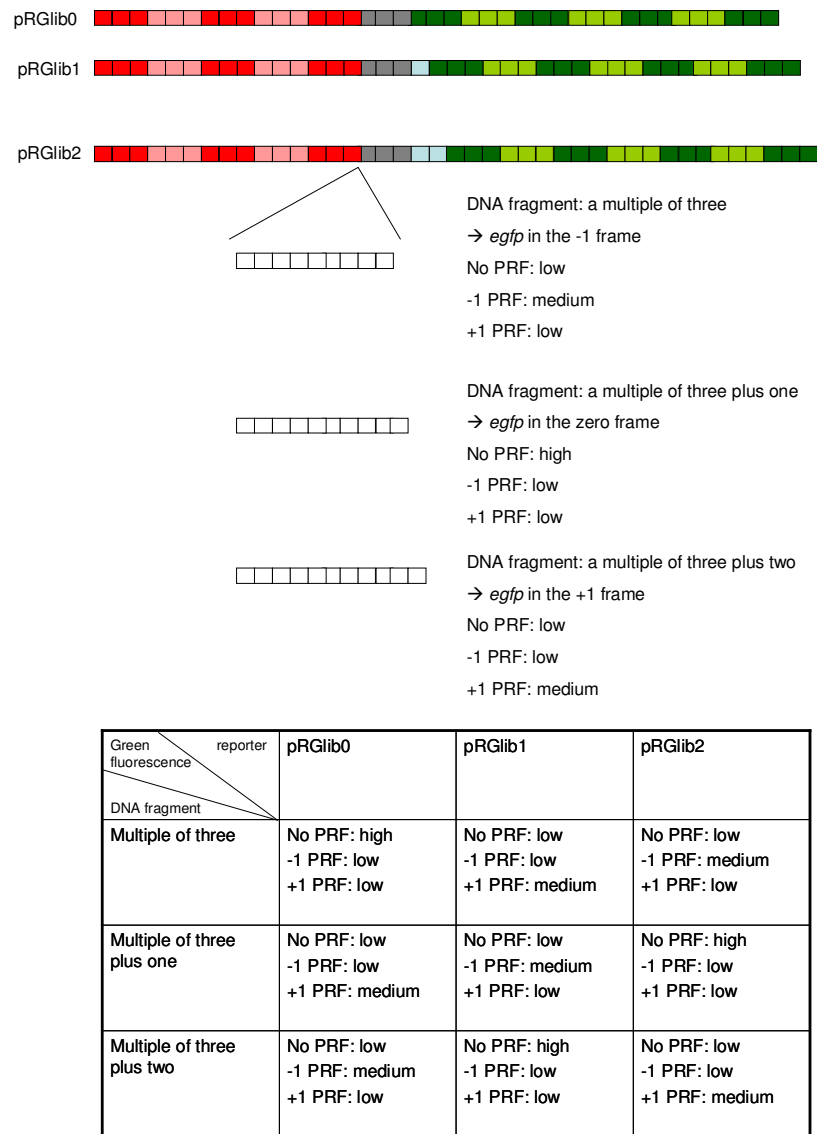


Figure 9.1 A reporter system for a genome-scale screening for programmed ribosomal frameshift (PRF) cassette. Top: representative sequence structures for three reporters, pRGlib0, pRGlib1, and pRGlib2. Red and pink boxes represent the coding sequence for DsRed. Green and light green boxes represent the coding sequence for EGFP. Gray boxes represent a stop codon in frame with DsRed. Light blue boxes represent nucleotide insertions for adjusting the reading frame of *egfp*. Using pRGlib2 as an example, cloning DNA fragments with various lengths would render *egfp* in different frames, which may result in different fluorescence patterns. A summary of the different green fluorescence patterns is listed at the bottom of the figure.



chemotactic response [7]. The present work has confirmed the presence of peptide derived from the PRF cassette in *yehP-egfp* fusion construct. The polypeptide sequences should be investigated for *pepP*, *nuoE*, and *cheA* to further confirm the frameshift site. Moreover, functional analysis of the frameshift products derived from these genes may provide more information regarding if +1 PRF in these genes are involved in gene regulation or protein function.

#### **9.2.4 Applying FSscan to other genomes**

In this dissertation, FSscan searched for +1 PRF hotspots in the *E. coli* genome. Because bacteriophage genes are translated by prokaryotic ribosomes, FSscan can potentially identify +1 PRF candidate in these genes. Additionally, the algorithm can be extended to other organisms with adjustments for the scoring system. In principle, the scoring system should account for stimulatory signals, tRNA: ribosome and tRNA: mRNA interactions, and the tRNA pool specific to the organism. Because the known +1 PRF cassettes are relatively diversified, caution needs to be taken for which organism-specific PRF features being incorporated into the program.

#### **9.2.5 Compositions of frameshift proteins in other -1 PRF cassettes**

This work developed mass spectrometry methods to differentiate two types of frameshift proteins generated from a -1 PRF cassette. As a proof of principle, the approaches were used to study PRF motifs from HIV-1 and bacteriophage P2 and PSP3. In addition, nano-liquid chromatography tandem mass spectrometry using multiple reaction monitoring revealed the change of the composition of the frameshift proteins when different steps in the translation cycle were perturbed. In a newly developed database Recode 2 [8], 257 genes are reported to involve -1 PRF. Whether these -1 PRF cassettes generate two types of frameshift proteins remain unclear. It is

possible that certain sequences trigger one pathway but suppress other pathways in the -1 PRF kinetic model. The investigation of other PRF motifs will advance our knowledge of how different sequence elements play their roles on -1 PRF in more detail.

### **9.3 Conclusion**

Programmed ribosomal frameshifting is an extension of genetic decoding that allows a ribosome to produce two types of polypeptides from a single mRNA. PRF occurs during translation elongation and involves several *cis*-acting elements. Computational and experimental tools developed in the present work allow analyzing this complex biological event from a systemic point of view. Kinetic models for +1 PRF and -1 PRF successfully explained experimental observations in the literature as well as identified critical steps in the mechanisms. Empirical examining model predictions in the dual fluorescence reporter system in *E. coli* demonstrated consistent results. The computational tools can be further adjusted to include more features in PRF. The dual fluorescence reporter system can be used for a large-scale anti-viral drug screening. The mass spectrometry method for the analysis of two types of frameshift proteins can be applied for other PRF signals to further understand the extent of different -1 PRF pathways. In conclusion, this dissertation work presents tools and strategies to effectively target critical steps in an important biological mechanism.

## REFERENCES

1. Plant,E.P. and Dinman,J.D. (2006) Comparative study of the effects of heptameric slippery site composition on -1 frameshifting among different eukaryotic systems. *RNA*, **12**, 666-673.
2. Cardno,T.S., Poole,E.S., Mathew,S.F., Graves,R. and Tate,W.P. (2009) A homogeneous cell-based bicistronic fluorescence assay for high-throughput identification of drugs that perturb viral gene recoding and read-through of nonsense stop codons. *RNA*, **15**, 1614-1621.
3. Dulude,D., Theberge-Julien,G., Brakier-Gingras,L. and Heveker,N. (2008) Selection of peptides interfering with a ribosomal frameshift in the human immunodeficiency virus type 1. *RNA*, **14**, 981-991.
4. Olsthoorn,R.C., Laurs,M., Sohet,F., Hilbers,C.W., Heus,H.A. and Pleij,C.W. (2004) Novel application of sRNA: Stimulation of ribosomal frameshifting. *RNA*, **10**, 1702-1703.
5. Henderson,C.M., Anderson,C.B. and Howard,M.T. (2006) Antisense-induced ribosomal frameshifting. *Nucleic Acids Res.*, **34**, 4302-4310.
6. Inoue,T., Shingaki,R., Hirose,S., Waki,K., Mori,H. and Fukui,K. (2007) Genome-wide screening of genes required for swarming motility in *Escherichia coli* K-12. *J. Bacteriol.*, **189**, 950-957.
7. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res* **28**:60-4.
8. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2009) Recode-2: New design, new search tools, and many more genes. *Nucleic Acids Res.*, doi:10.1093/nar/gkp788.